ALESSANDRA LOUZADA TERRA

APRENDIZADO DE MÁQUINA E ANÁLISE DE REDES SOCIAIS EM GRAFOS: UM ESTUDO DO MOVIMENTO DOS ESTUDANTES DE MEDICINA NO BRASIL

LAVRAS – MG

2025

ALESSANDRA LOUZADA TERRA

APRENDIZADO DE MÁQUINA E ANÁLISE DE REDES SOCIAIS EM GRAFOS:
UM ESTUDO DO MOVIMENTO DOS ESTUDANTES DE MEDICINA NO BRASIL

Dissertação apresentada à Universidade Federal de Lavras, como parte das exigências do Programa de Pós-Graduação em Ciência da Computação para obtenção do título de Mestre.

Prof. DSc. Eric Fernandes de Mello Araújo Orientador

LAVRAS – MG

2025

Ficha catalográfica elaborada pela Coordenadoria de Processos Técnicos da Biblioteca Universitária da UFLA

Terra, Alessandra Louzada
Aprendizado de máquina e análise de redes sociais em grafos: um estudo do movimento dos estudantes de medicina no Brasil / Alessandra Louzada Terra. – Lavras : UFLA, 2025.
67 p. :

Dissertação (mestrado)–Universidade Federal de Lavras, 2025.
Orientador: Prof. DSc. Eric Fernandes de Mello Araújo.
Bibliografia.

1. Aprendizado de Máquina. 2. Redes Sociais. 3. Mobilidade Médica. I. Universidade Federal de Lavras. II. Título.

ALESSANDRA LOUZADA TERRA

APRENDIZADO DE MÁQUINA E ANÁLISE DE REDES SOCIAIS EM GRAFOS: UM ESTUDO DO MOVIMENTO DOS ESTUDANTES DE MEDICINA NO BRASIL

MACHINE LEARNING AND SOCIAL NETWORK ANALYSIS IN GRAPHS: A STUDY OF THE MEDICAL STUDENT MOVEMENT IN BRAZIL

Dissertação apresentada à Universidade Federal de Lavras, como parte das exigências do Programa de Pós-Graduação em Ciência da Computação para obtenção do título de Mestre.

APROVADA em 30 de janeiro de 2025.

Prof. DSc. Michele Amaral Brandão	UFMG
Prof. DSc. Marluce Rodrigues Pereira	UFLA
Prof. DSc. Mayron Cesar de Oliveira Moreira	UFLA

Prof. DSc. Eric Fernandes de Mello Araújo Orientador

LAVRAS – MG

2025

LISTA DE FIGURAS

LISTA DE TABELAS

LISTA DE QUADROS

RESUMO

O sistema de saúde do Brasil vem enfrentando dificuldades há décadas devido à distribuição desigual de médicos no país. Poucos estudos tentaram abordar a mobilidade dos médicos para entender quais são os fatores de decisão que determinam onde os profissionais se estabelecerão. O conhecimento sobre os padrões de circulação dos médicos no Brasil pode ser de grande valia para o governo, pois fornecerá informações que podem levar a melhores políticas de oportunidades de trabalho, bem como definir melhores locais para novas escolas médicas. Mais especificamente, é importante entender como os estudantes de medicina decidem para onde ir para a graduação, pois isso afetará a mobilidade dos profissionais no futuro. Este trabalho é parte de uma investigação de como é o fluxo de médicos no Brasil levando em consideração os dados fornecidos pelo Ministério da Saúde e outras agências de pesquisa e governamentais brasileiras. O estudo aqui proposto utiliza técnicas de aprendizado de máquina para derivar e entender os padrões de onde as pessoas se formam e exercem a profissão. Também é utilizada a Análise de Redes Sociais a fim de avaliar os maiores fluxos de pessoas e a migração de pessoas entre as regiões. Os resultados revelam que estados com maior centralidade na rede tendem a atrair mais médicos, tanto na fase de formação quanto no exercício profissional. Além disso, identificamos correlações entre o local de graduação e a escolha do local de trabalho, evidenciando que a mobilidade médica segue padrões concentrados em determinadas regiões. Essas descobertas podem contribuir para políticas de incentivo que favoreçam a redistribuição equitativa de médicos pelo país.

Palavras-chave: médicos; aprendizado de máquina; previsões; mobilidade de estudantes.

ABSTRACT

Brazil's healthcare system has been facing challenges for decades due to the uneven distribution of doctors across the country. Few studies have attempted to address medical mobility to understand the decision-making factors that determine where professionals choose to settle. Understanding the circulation patterns of doctors in Brazil can be highly valuable for the government, as it provides insights that may lead to better job opportunity policies and help define optimal locations for new medical schools. More specifically, it is crucial to understand how medical students decide where to pursue their degrees, as this choice will influence the future mobility of professionals. This study is part of an investigation into the flow of doctors in Brazil, considering data provided by the Ministry of Health and other Brazilian research and governmental agencies. The proposed study employs machine learning techniques to derive and analyze patterns in where individuals graduate and practice medicine. Additionally, Social Network Analysis is used to assess the major migration flows of medical professionals across regions. The results indicate that states with higher centrality in the network tend to attract more doctors, both during their education and in their professional practice. Furthermore, we identified correlations between the location of medical education and the choice of workplace, highlighting that medical mobility follows concentrated patterns in specific regions. These findings can contribute to incentive policies that promote a more equitable redistribution of doctors throughout the country.

Keywords: physicians; machine learning; predictions; students' mobility**.**

INDICADORES DE IMPACTO

Esta pesquisa apresenta impactos multidisciplinares significativos ao investigar a mobilidade de médicos no Brasil através de técnicas de ciência de dados. Socialmente, o estudo revela padrões de desigualdade na distribuição de profissionais de saúde, identificando regiões com carência de médicos que demandam políticas públicas específicas, impactando diretamente a qualidade do atendimento à população. Tecnologicamente, desenvolve metodologias inovadoras ao aplicar algoritmos de aprendizado de máquina e análise de redes sociais para mapear fluxos migratórios, oferecendo ferramentas para gestão estratégica de recursos humanos em saúde. Economicamente, os resultados permitem otimizar investimentos em educação médica e programas de fixação de profissionais, potencialmente reduzindo custos com a rotatividade e melhorando a eficiência do sistema de saúde. Culturalmente, o trabalho contribui para compreender como fatores regionais e educacionais influenciam as escolhas profissionais, revelando padrões de mobilidade que refletem desigualdades estruturais no país. O estudo impacta principalmente as áreas temáticas de Saúde (6) e Trabalho (8), alinhando-se aos Objetivos de Desenvolvimento Sustentável 3 (Saúde e Bem-estar) e 10 (Redução das Desigualdades) da ONU. Os resultados beneficiam gestores públicos, instituições de ensino médico e formuladores de políticas de saúde em todo território nacional, com potencial para influenciar a distribuição de mais de 500 mil profissionais da saúde no país. A pesquisa estabelece parcerias implícitas com órgãos públicos de saúde através da utilização de dados oficiais, embora não envolva diretamente ações extensionistas.

IMPACT INDICATORS

This research presents significant multidisciplinary impacts by investigating physician mobility in Brazil through data science techniques. Socially, the study reveals patterns of inequality in health professional distribution, identifying regions with physician shortages that require specific public policies, directly affecting population healthcare quality. Technologically, it develops innovative methodologies by applying machine learning algorithms and social network analysis to map migration flows, offering tools for strategic human resource management in healthcare. Economically, the results allow optimization of investments in medical education and professional retention programs, potentially reducing turnover costs and improving health system efficiency. Culturally, the work contributes to

understanding how regional and educational factors influence professional choices, revealing mobility patterns that reflect structural inequalities in the country. The study primarily impacts the thematic areas of Health (6) and Work (8), aligning with UN Sustainable Development Goals 3 (Good Health and Well-being) and 10 (Reduced Inequalities). The results benefit public managers, medical education institutions, and health policy makers nationwide, with potential to influence the distribution of over 500,000 healthcare professionals in the country. The research establishes implicit partnerships with public health agencies through the use of official data, though it does not directly involve extension activities.

SUMÁRIO

1 INTRODUÇÃO

Num país de dimensão continental, como o Brasil, a mobilidade dos médicos tem um grande impacto na distribuição dos recursos humanos no sistema de saúde em todo o país. Cidades maiores e com melhores condições de trabalho atraem mais médicos, criando uma maior concentração de recursos humanos em determinadas regiões e uma falta em outras partes do país. O primeiro passo para ser médico é a escolha de onde estudar. E esta decisão pode afetar a localização e as oportunidades que os médicos formados podem encontrar em seu futuro. Existem atualmente 323 escolas de medicina que oferecem cerca de 32.000 vagas no Brasil. A maior parte das escolas está na região Sudeste (42%), enquanto apenas 6,8% das escolas estão localizadas na região Norte. A distribuição das escolas é um fator importante, pois a maioria das pessoas são atraídas para determinadas regiões mais desenvolvidas, enquanto as pessoas de regiões mais pobres têm menos oportunidades de estudar medicina (OLIVEIRA et al., 2019).

A escassez de escolas em certas regiões é um fator significativo na migração de estudantes dessas áreas para as regiões mais prósperas do país. As pessoas tendem a não retornar aos seus lugares de origem em busca de uma carreira, já que os novos locais proporcionam melhores rendimentos e condições de trabalho. Por esta razão, é relevante que o governo e outras partes interessadas no sistema de saúde sejam informados sobre o impacto da geolocalização e outras características que contribuem para determinar onde os médicos formados irão fixar suas residências. A hipótese é que o conhecimento de como funciona a mobilidade dos estudantes seja fundamental no planejamento da abertura de novas escolas e na detecção de problemas estruturais que possam ser a causa da falta de especialistas em certas regiões.

Este trabalho faz parte de um estudo mais amplo sobre a circularidade dos médicos, também conhecido como mobilidade médica. A circularidade médica diz respeito ao fluxo de médicos em relação à localização geográfica dos locais de trabalho. Este tema é muito relevante devido à crônica situação brasileira de falta de médicos em regiões específicas, geralmente com Índice de Desenvolvimento Humano (IDH) mais baixo ou regiões mais isoladas. Este fenômeno também está presente em outros países e pouco se sabe sobre os motivos pelos quais ocorre essa desigualdade na distribuição de recursos (ANAND et al., 2008; MAKINEN et al., 2000; RóJ, 2020).

Os avanços na área de Ciência de Dados, combinados com a disponibilidade de dados

confiáveis dos governos, abre novas oportunidades para investigar as razões pelas quais as pessoas preferem determinados lugares, mas também fornece material para simular cenários e prever políticas potenciais para mitigar o problema. Os algoritmos utilizados nesse trabalho consideram a localização de origem e outras características das pessoas, como índice de desenvolvimento humano. Ao mapear o fluxo de alunos desde o nascimento até a universidade, essa previsão ajuda a identificar a relação entre o local de nascimento, o local de estudo e o local onde o trabalho médico será realizado. Além do teste dos algoritmos, foram realizadas análises de redes sociais utilizando teoria de grafos para entender melhor os padrões de migração ao longo do tempo. Essas análises permitiram identificar e visualizar as conexões e interações entre diferentes regiões, destacando como e quando os estudantes se deslocam de um local para outro. Ao examinar esses padrões em uma escala temporal, conseguimos obter informação sobre as tendências de migração, identificar possíveis fatores de influência e compreender melhor estas dinâmicas. Os conjuntos de dados utilizados para a análise foram fornecidos pelo Sistema Unico de Saúde (SUS) e Instituto Nacional de Estudos e Pesquisas (INEP).

O objetivo geral deste estudo foi estudar e analisar os dados da mobilidade de estudantes de medicina por meio de grafos aplicando algoritmos de aprendizado de máquina para fins de identificação de padrões de escolhas de escolas de medicina a partir de uma base de dados sobre estudantes de medicina. Para alcançar esse propósito, os objetivos específicos incluíram:

a) a análise detalhada da movimentação dos estudantes por meio de estudos das redes sociais, com foco em comparações de centralidade para identificar pontos-chave de interação;
b) a utilização de grafos para realizar as análises de redes sociais, facilitando a compreensão dos fluxos de movimentação e dos relacionamentos entre os diversos locais de estudo e atuação dos estudantes de medicina no Brasil;
c) a utilização de dois diferentes tipos de algoritmo de classificação (Rede Neural Artificial e *Random Forest*), para analisar e avaliar seus desempenhos em termos de acurácia, precisão, *recall* e *F1-score*. A análise busca identificar qual algoritmo oferece melhores resultados na previsão dos padrões de mobilidade estudantil, contribuindo para a compreensão dos fatores que influenciam a escolha do local de graduação em medicina;
d) o uso de gráficos, mapas de calor e coropléticos, para ilustrar a movimentação dos indivíduos.

Com isso, espera-se identificar tendências e fatores que influenciam a escolha dos locais de formação e trabalho desses profissionais. Os resultados obtidos mostram que as técnicas de Aprendizado de Máquina e Análise de Redes Sociais foram eficazes para prever o local de estudos dos estudantes de medicina e analisar a mobilidade dos profissionais na área médica. Os modelos de Rede Neural Artificial e *Random Forest* tiveram bom desempenho, com uma acurácia superior a 90%, com o *Random Forest* se destacando por sua maior precisão. A análise de centralidade identificou padrões regionais de mobilidade. Mapas de calor e coropléticos revelaram uma democratização do acesso ao ensino superior e mostraram que muitos profissionais tendem a permanecer próximos ao local onde estudaram. Esses resultados fornecem subsídios para políticas públicas voltadas à distribuição equilibrada de oportunidades educacionais e à retenção de profissionais em áreas menos atendidas.

Esta dissertação está organizada em seis capítulos. Este primeiro capítulo apresenta a introdução, contextualizando o problema, os objetivos do estudo e a relevância da pesquisa. O segundo capítulo aborda os trabalhos relacionados, discutindo estudos prévios sobre mobilidade médica e técnicas de análise utilizadas na literatura. No terceiro capítulo, são explorados os conceitos fundamentais no referencial teórico, incluindo aprendizado de máquina e análise de redes sociais. O quarto capítulo descreve a metodologia, detalhando os dados utilizados, os algoritmos empregados e os procedimentos adotados para a análise. No quinto capítulo, são apresentados os resultados e discussões, onde os principais achados da pesquisa são analisados e interpretados. Por fim, o sexto capítulo traz a conclusão final, sintetizando as contribuições do estudo, suas limitações e sugestões para trabalhos futuros.

2 TRABALHOS RELACIONADOS

Os estudos sobre mobilidade e migração concentram-se principalmente na geopolítica e na demografia, visando encontrar os fatores que influenciam as pessoas a mudarem de localização (OKóLSKI, 2012; CELBIs¸, 2021). Até onde foi possível investigarmos, não existem trabalhos utilizando aprendizado de máquina para investigar especificamente a migração de estudantes de medicina. Abaixo, são apresentados os trabalhos sobre mobilidade médica e sobre mobilidade em geral que utilizam técnicas de aprendizado de máquina.

Wang et al. (2019) usam dados heterogêneos provenientes de terminais portáteis, GPS e mídias sociais, para explorar os padrões de mobilidade humana urbana de forma quantitativa e microscópica. Nesse levantamento são revisados modelos de mobilidade humana sob uma perspectiva centrada no ser humano em um contexto orientado por dados. Para isso, foram examinadas as características e métricas dos padrões de mobilidade humana, destacando aspectos espaciais, temporais e sociais. Os padrões de mobilidade humana são caracterizados em níveis individual, coletivo e híbrido, enquanto os métodos de previsão são examinados sob quatro aspectos, seguidos de uma descrição dos desenvolvimentos recentes. Em seguida, foram exploradas diversas técnicas de previsão de mobilidade, agrupadas em quatro categorias principais: baseadas em Markov, baseadas em compressão, baseadas em séries temporais e baseadas em aprendizado de máquina. Por fim, são discutidas questões em aberto que fornecem uma referência útil para a direção futura dos pesquisadores. Este estudo não apenas estabelece uma base sólida para iniciantes que desejam compreender rapidamente a mobilidade humana, mas também fornece informações úteis para pesquisadores sobre como desenvolver um modelo unificado de mobilidade humana.

Asgari, Gauthier and Becker (2013) apresentam uma revisão abrangente dos modelos e métodos de previsão de mobilidade humana em ambientes urbanos, com ênfase na análise de dados e comportamentos humanos. O trabalho aborda características e métricas dos padrões de mobilidade humana, explorando desde estudos baseados em trajetórias até estudos que utilizam teoria de grafos e redes. Ao analisar medidas estatísticas em estudos baseados em trajetórias, como distribuição de comprimentos de salto e raio de giração, busca-se investigar como as pessoas se movem em seu cotidiano e se é possível modelar esses movimentos individuais para fazer previsões com base neles. A utilização de grafos nos estudos de mobilidade ajuda a investigar o comportamento dinâmico do sistema, como difusão e fluxo na rede, facilitando a estimativa de quanto uma parte da rede influencia outra por meio de métricas como medidas de centralidade. O estudo tem como objetivo estudar o fluxo populacional em redes de transporte

usando dados de mobilidade para derivar modelos e padrões e desenvolver novas aplicações na previsão de fenômenos como congestionamento. No entanto, com o aumento no volume de dados gerado pela conectividade atual, especialmente por redes de telefonia celular, surgem novos desafios relacionados ao armazenamento, à representação, à análise e à complexidade computacional desses dados.

Feng et al. (2018) abordam a predição de mobilidade humana. Identificam-se três desafios principais na predição da mobilidade: a complexidade das regularidades sequenciais de transição, a periodicidade multi-nível da mobilidade humana e a heterogeneidade e a escassez dos dados de trajetória. Diante desses desafios, é proposto o *DeepMove*, uma rede neural recorrente com atenção para a predição da mobilidade a partir de trajetórias extensas e esparsas. O DeepMove combina um modelo de rede neural recorrente com uma atenção histórica para capturar tanto as transições sequenciais complexas quanto a periodicidade multi-nível da mobilidade humana. Os resultados experimentais demonstram que o valor da acurácia do DeepMove supera em mais de 10% os modelos de predição de mobilidade do estado-da-arte.

Seixas et al. (2019) analisam a movimentação dos médicos em cinco regiões do Estado de São Paulo, Brasil, sob uma nova perspectiva denominada “circularidade médica”. Essa abordagem destaca a diversidade de vínculos profissionais dos médicos ao longo do tempo e em diferentes espaços geográficos. Utilizando uma metodologia de estudo de casos múltiplos, combinando abordagens quantitativas e qualitativas, os pesquisadores categorizaram os médicos como “exclusivos”, com vínculos apenas na região em análise, e “não exclusivos”, com vínculos em outras regiões também. Os resultados revelaram uma dependência regional de médicos externos variando de 30% a 40%, sendo mais elevada em regiões mais desenvolvidas e menor nas menos desenvolvidas. Médicos não exclusivos foram identificados como mais especializados, especialmente em áreas cirúrgicas e de diagnóstico, enquanto os exclusivos tendem a atuar em especialidades básicas e clínicas. Esta pesquisa destaca a importância de compreender a dinâmica da força de trabalho médica em nível regional para informar políticas de redistribuição mais eficazes e integradas no setor de saúde.

Seixas et al. (2017) visam caracterizar a circularidade médica no Brasil com foco nas regiões Norte-Barretos e Sul-Barretos, em São Paulo. Utilizando um método transversal e análise de dados secundários, constatou-se que, em média, 45% dos médicos em atividade circulam entre diferentes regiões de saúde no país, com mais de 50% da força de trabalho médica atuando em áreas distintas de suas regiões originais. Os profissionais com maior propensão à mobilidade são aqueles especializados em cirurgia e serviços de apoio ao

diagnóstico e terapia. O trabalho conclui que a alta circulação de médicos entre regiões possui características específicas dependendo da localização geográfica e da estrutura de saúde. No caso de Barretos, a movimentação dos profissionais está associada à sua especialização e afiliação profissional. Esses achados fornecem *insights* importantes para entender a dinâmica da força de trabalho médica e podem auxiliar na formulação de políticas públicas mais eficazes para o setor de saúde. Seixas et al. (2017) visam caracterizar a circularidade médica no Brasil com foco nas regiões Norte-Barretos e Sul-Barretos, em São Paulo. Utilizando um método transversal e análise de dados secundários, constatou-se que, em média, 45% dos médicos em atividade circulam entre diferentes regiões de saúde no país, com mais de 50% da força de trabalho médica atuando em áreas distintas de suas regiões originais. Os profissionais com maior propensão à mobilidade são aqueles especializados em cirurgia e serviços de apoio ao diagnóstico e terapia. O trabalho conclui que a alta circulação de médicos entre regiões possui características específicas dependendo da localização geográfica e da estrutura de saúde. No caso de Barretos, a movimentação dos profissionais está associada à sua especialização e afiliação profissional. Esses achados fornecem *insights* importantes para entender a dinâmica da força de trabalho médica e podem auxiliar na formulação de políticas públicas mais eficazes para o setor de saúde.

Costigliola (2011) traz uma compilação dos principais fatores que contribuem para a migração de médicos na Europa e discute as atuais tendências e lacunas na mobilidade desses profissionais no continente. O estudo mostra que na Europa esse movimento é observado desde a década de 1940 e, com a integração europeia, surgiram novas oportunidades para os profissionais se formarem, estudarem ou trabalharem melhor em outros países. Aponta-se que essa migração ocorre devido a problemas do sistema de saúde, onde os médicos se deslocam para outras localidades em busca de melhores condições de trabalho, oportunidades e reconhecimento social. O artigo aponta que a circularidade médica apresenta vantagens e desvantagens, onde o benefício é quando a movimentação ocorre temporariamente, para estudar e vivenciar, e o profissional retorna ao local de origem. A desvantagem é quando essa mudança ocorre em longo prazo, gerando problemas no sistema de saúde. Por fim, o estudo conclui que a migração de médicos gera impactos tanto no país de origem quanto no país de destino, além de apontar como fatores para o movimento, o impacto das crises econômicas, a incerteza da evolução da força de trabalho em saúde na Europa e a contratação de profissionais de saúde de outros países para compensar a escassez em determinadas regiões.

No Quadro 2.1, é possível visualizar alguns trabalhos que abordam separadamente a mobilidade médica e a ciência de dados, sem uma interseção direta entre os dois temas.

Quadro 2.1 - Trabalhos relacionados sobre mobilidade médica e ciência de dados.

#	**Trabalhos Relacionados**	**Mobilidade Médica**	**Ciência de Dados**
1	Wang et al. (2019), Urban human mobility: Data-driven modeling and prediction.	×	✓
2	Asgari, Gauthier and Becker (2013), A survey on human mobility and its applications.	×	✓
3	Feng et al. (2018), Deepmove: Predicting human mobility with attentional recurrent networks	×	✓
4	Seixas et al. (2019), A circularidade dos médicos em cinco regiões de São Paulo Brasil: padrões e fatores intervenientes	✓	×
5	Seixas et al. (2017), Physicians' circularity in health regions in Brazil.	✓	×
6	Costigliola (2011), Mobility of medical doctors in cross-border healthcare.	✓	×

Fonte: Da autora (2025).

Como visto é possível perceber que o aprendizado de máquina pode ser aplicado aos mais variados tipos de estudo, com diferentes abordagens e técnicas, assim como os estudos já realizados para entender a mobilidade médica foram feitos de forma manual através de dados e resultados das áreas competentes do governo. Portanto, o trabalho é relevante porque propõe o uso de técnicas de aprendizado de máquina, grafos e análise de redes sociais para entender o funcionamento e a movimentação dos estudantes de medicina, bem como prever onde possíveis futuros alunos poderão vir estudar. Pode, portanto, auxiliar na identificação da relação entre o local de nascimento, o local de estudo e o possível local de trabalho do recém-formado.

3 REFERENCIAL TEÓRICO

Neste capítulo apresentamos uma revisão de literatura e enfatizamos a escassez de trabalhos utilizando aprendizado de máquina para investigar, especificamente, a migração de estudantes de medicina. Esta escassez de trabalhos foi o que motivou este estudo. Também explicaremos as técnicas e os conceitos a serem utilizados para o desenvolvimento das análises.

3.1 Mobilidade Médica

A mobilidade de médicos pode ser entendida como a movimentação de médicos que ocorre de forma mais duradoura onde o indivíduo se desloca para outro local e se instala naquele local de forma permanente (SEIXAS et al., 2019). O tema vem sendo estudado desde a década de 1940 no contexto da migração de médicos em território europeu (SEIXAS et al., 2017). Estudos têm tentado abordar a questão do por quê médicos decidem se mudar para outras localidades, atribuindo a decisão a fatores como as condições do sistema de saúde e a atratividade das localidades (CAMPOS; MALIK, 2008; COSTIGLIOLA, 2011).

Um estudo canadense, (BERGEVIN1 et al., 2015), mostrou que as condições do sistema de saúde e as características do local são relevantes na hora de os médicos decidirem onde trabalhar. O estudo enfatiza que a infraestrutura e a eficiência dos sistemas regionais de saúde impactam na escolha dos médicos quanto ao local de trabalho. Fatores como integração dos serviços, disponibilidade de recursos, apoio à saúde comunitária e condições de governança desempenham um papel importante na decisão dos profissionais. Regiões que oferecem melhores condições para a prática médica e acesso a tecnologia, além de um ambiente bem coordenado entre hospitais e cuidados primários, tornam-se mais atraentes para médicos. No entanto, áreas rurais e menos desenvolvidas enfrentam dificuldades na retenção de profissionais devido à falta de estrutura adequada e limitações no suporte. A inclusão de sistemas de remuneração baseados em desempenho e incentivos financeiros também são discutidos como estratégias para aumentar a atratividade dessas regiões. Essas dinâmicas sugerem que a transformação dos sistemas de saúde pode influenciar diretamente a distribuição e retenção de profissionais, ajudando a melhorar o equilíbrio regional no acesso aos cuidados de saúde.

Esperamos com nossa pesquisa avançar o entendimento da mobilidade de médicos por meio da aplicação de aprendizado de máquina e análise de redes sociais em grafos. Essa

abordagem busca identificar padrões a partir de dados que conectam a origem e o destino dos estudantes de graduação em medicina. Até agora, não encontramos essa metodologia na literatura existente, o que destaca a originalidade e a importância do nosso trabalho para explorar as dinâmicas de mobilidade na área da saúde.

3.2 Aprendizado de Máquina

O aprendizado de máquina, ou *Machine Learning*, pode ser entendido como um campo de estudo da Inteligência Artificial cujo objetivo é "o desenvolvimento de técnicas computacionais de aprendizado, bem como a construção de sistemas capazes de adquirir conhecimento automaticamente" (MONARD; BARANAUSKAS, 2003). O termo foi cunhado por Arthur Samuel em 1959, que descreveu a área então recentemente criada como "um campo de estudo que dá aos computadores a capacidade de aprender sem serem programados para isso" (GERON, 2019). Mesmo tendo sido criado na virada da década de 1950, algoritmos de aprendizado de máquina só se tornaram realmente úteis a partir da década de 1990, com o crescimento do uso da internet, à medida que o número de dados a serem processados aumentou significativamente (GERON, 2019).

Existem vários tipos de técnicas de aprendizado de máquina, podendo ser divididas em três principais grupos: (1) Aprendizado supervisionado; (2) Aprendizado não-supervisionada e; (3) Aprendizado por reforço.

Algoritmos de aprendizado supervisionado utilizam conjuntos de dados rotulados para o treinamento, o que significa que cada exemplo no conjunto de dados inclui a entrada (características ou atributos) e a saída desejada (rótulo ou valor alvo) (BODO, 2015). O conjunto de dados utilizado neste trabalho é adequado para este tipo de técnica, pois possui todas as características dos indivíduos e também o local onde foram estudar medicina. Assim, temos um conjunto de dados rotulado.

Algoritmos de aprendizado não-supervisionados fazem uso de um conjunto de dados não-rotulados. Ou seja, o algoritmo não recebe as informações do objetivo a ser encontrado. Em vez disso, esses tipos de algoritmos podem utilizar da técnica de agrupar os dados de acordo com suas características (traços) para encontrar semelhanças entre eles. Após o modelo ser treinado com o conjunto de dados, ele pode ser usado para classificar novos dados de entrada, permitindo estudos e descobertas sobre a estrutura dos dados sem a necessidade de rótulos pré-existentes. (BODO, 2015).

Algoritmos de aprendizado por reforço são uma mistura de algoritmos de aprendizado não-supervisionado e algoritmos de aprendizado supervisionado, onde um agente interage com um ambiente dinâmico tomando ações e recebendo resposta na forma de recompensas ou penalidades. O agente tem como objetivo aprender a escolher sequências de ações que maximizem a recompensa acumulada ao longo do tempo. No aprendizado por reforço não há uma correspondência direta entre entradas e saídas. Ao invés disso, o agente aprende por tentativa e erro, ajustando suas ações de acordo com as recompensas recebidas do ambiente. Geralmente, não é preciso fornecer rótulos para todo o conjunto de dados ao criar um modelo de aprendizado por reforço. Em vez de isso, usam-se algumas amostras de dados rotulados para definir automaticamente as recompensas que o agente tentará maximizar. A preparação dos dados é reduzida, uma vez que o agente aprende diretamente com o ambiente, em vez de depender de um grande conjunto de dados previamente rotulado (BODO, 2015).

Neste trabalho, nosso objetivo foi de utilizar algoritmos de aprendizado supervisionado para prever o destino dos estudantes de medicina no Brasil. Para isso, testamos dois algoritmos: *Random Forest* e Redes Neurais, que serão detalhados nas subseções seguintes. Além disso, nossa análise inclui uma investigação dos aspectos de rede relacionados à mobilidade dos médicos. Para essa parte do estudo, aplicamos a análise de redes sociais e a teoria de grafos, visando entender melhor as dinâmicas e os padrões de deslocamento.

3.2.1 Random Forest

A escolha para utilização do algoritmo *Random Forest* se deu devido a sua robustez e eficácia na análise de conjuntos de dados complexos, como o utilizado nesse trabalho (RIGATTI, 2017), onde o o autor explora o uso da técnica *Random Forest* na análise de dados de sobrevivência, especialmente para identificar fatores de risco associados à mortalidade. Este método é menos suscetível ao *overfitting*[1], o que contribui para a precisão das previsões. Além disso, a capacidade do *Random*
Forest de fornecer importâncias de características (RIGATTI, 2017) permite identificar os fatores que mais influenciam a mobilidade, oferecendo *insights* valiosos que podem apoiar a formulação de políticas na área da saúde.

Para facilitar o entendimento do *Random Forest* vamos primeiramente entender como funciona o método Árvore de Decisão (SMITH, 2017). Árvore de Decisão é um algoritmo de aprendizado de máquina supervisionado que apresenta um conjunto de regras de decisão organizadas em uma estrutura de árvore, na qual os nós internos da árvore representam

critérios de decisão baseados em diferentes atributos dos dados. Cada nó divide os dados em subconjuntos mais homogêneos, levando em consideração a classe de saída, o que é essencial para a classificação. É importante controlar a profundidade da árvore, já que árvores muito profundas podem levar ao problema de *overfitting*, enquanto árvores muito rasas podem subestimar a complexidade dos dados (MONARD; BARANAUSKAS, 2003).

O algoritmo *Random Forest*, por sua vez, gera uma floresta aleatoriamente a partir de uma combinação de árvores de decisão. Ao criar essas árvores, o método adiciona uma camada extra de aleatoriedade ao modelo, selecionando aleatoriamente um subconjunto de características em cada nó de divisão. Dentro desse subconjunto, a melhor característica é escolhida para dividir os dados. A ideia é criar uma grande diversidade de possibilidades de classificação, o que geralmente leva a uma melhor geração de modelos e o torna uma boa opção para uso em dados robustos. Em outras palavras, *Random Forest* cria diversas árvores de decisão e as combina para obter uma previsão com maior precisão e mais estabilidade.

Dentro do algoritmo *Random Forest*, existem dois parâmetros importantes: o a profundidade máxima (*max_depth*) e a quantidade de árvores(*n_estimators*). O parâmetro *max_depth* controla a profundidade máxima das árvores de decisão individuais na floresta. Ao limitar a profundidade das árvores, é possível evitar o *overfitting*. Definir um valor para *max_depth* impõe uma restrição à complexidade das árvores, prevenindo que elas se tornem muito profundas e especializadas nos dados de treinamento. Por outro lado, *n_estimators* determina o número de árvores de decisão na floresta. Quanto maior o número de estimadores, mais robusto e poderoso será o modelo, pois terá mais árvores para fazer previsões e mitigar o impacto de variações nos dados de entrada. No entanto, adicionar mais árvores também aumenta o custo computacional do treinamento e da previsão, sendo necessário encontrar um equilíbrio entre desempenho e eficiência computacional.

Para aos valores desses parâmetros do modelo, a técnica de busca em grade (*grid-search*) pode ser empregada. O *grid-search* consiste em um método sistemático de hiperparametrização que realiza uma busca exaustiva entre os valores especificados para cada hiperparâmetro. Por meio dessa técnica, é criada uma grade de parâmetros na qual o modelo é treinado repetidamente, testando todas as combinações possíveis dos parametros.

[1] Overfitting ocorre quando um modelo de aprendizado de máquina se ajusta excessivamente aos dados de treinamento, capturando tanto padrões gerais quanto ruídos ou detalhes irrelevantes.

3.2.2 Redes Neurais

Redes neurais são algoritmos de aprendizado de máquina que se inspiram no funcionamento do cérebro humano e são particularmente eficazes em tarefas complexas, como reconhecimento de voz, processamento de linguagem natural e análise de imagens. Sua estrutura é composta por neurônios organizados em camadas: camada de entrada, camadas ocultas que realizam transformações e camada de saída. O processo de treinamento inclui etapas de inicialização, propagação direta, cálculo da perda, retro-propagação, otimização e validação/teste. Essas etapas são comuns a todas as redes neurais, embora os detalhes possam variar conforme a arquitetura utilizada (GOODFELLOW; BENGIO; COURVILLE, 2016).

O processo de treinamento de uma rede neural envolve várias etapas. Primeiro, os pesos dos neurônios são inicializados aleatoriamente. Em seguida, ocorre a propagação direta, onde os dados são passados pela rede, camada por camada, produzindo uma saída predita. Após isso, é calculada uma função de perda que mede a diferença entre as previsões da rede e os rótulos reais dos dados de treinamento. Em seguida, a retro-propagação é usada para calcular os gradientes da função de perda em relação aos pesos da rede, permitindo ajustá-los para reduzir a perda. Finalmente, o desempenho da rede é avaliado em um conjunto de dados de validação ou teste para garantir que ela esteja generalizando bem para dados não vistos.

Existem várias arquiteturas de redes neurais, cada uma projetada para tarefas específicas. As redes profundas, por exemplo, possuem várias camadas ocultas, permitindo-as aprender representações complexas e hierárquicas dos dados. As redes convolucionais (CNNs) são amplamente utilizadas em tarefas de visão computacional, devido à sua capacidade de capturar padrões espaciais nos dados de imagem de forma eficiente. Por outro lado, as redes recorrentes (RNNs) são adequadas para lidar com dados sequenciais ou temporais, como séries temporais ou linguagem natural, graças à sua capacidade de processar sequências de comprimento variável e capturar dependências temporais. Essas diferentes arquiteturas de redes neurais oferecem flexibilidade para lidar com uma ampla gama de problemas de aprendizado de máquina (IBM, 2022).

O *TensorFlow*[2] é uma ferramenta utilizada para a implementação dessas estruturas. *TensorFlow* é uma biblioteca de código aberto desenvolvida pelo Google, projetada especificamente para construir e treinar redes neurais e outros modelos de aprendizado de máquina. Com o *TensorFlow*, é possível criar facilmente a estrutura da rede neural, definindo as camadas

[2] https://www.tensorflow.org/?hl=pt-br

unidades e conexões entre elas. Ele oferece uma interface flexível e de alto nível que simplifica o processo de desenvolvimento, permitindo que os usuários se concentrem mais na lógica do modelo do que em detalhes de implementação de baixo nível (TENSORFLOW, 2024).

Dentre os parâmetros para a compilação do modelo com o *TensorFlow*, temos as funções de perda “sparse_categorical_crossentropy” e “categorical_crossentropy” e os otimizadores “adam” e “SGD (Stochastic Gradient Descent)”. Neste trabalho utilizamos o “adam” e a “sparse_categorical_crossentropy”

O otimizador “adam” é utilizado para ajustar os pesos de uma rede neural durante o treinamento com o objetivo de minimizar a função de perda. Ele combina os conceitos de “momentum” (momento) e “RMSprop” (média dos quadrados dos gradientes) para atualizar os pesos de forma eficiente. O otimizador adapta a taxa de aprendizado de forma dinâmica para cada parâmetro, ajustando-a com base na média dos gradientes e na variância dos gradientes recentes (VISHWAKARMA, 2024).

A função de perda “sparse_categorical_crossentropy” é uma medida que quantifica a diferença entre as probabilidades previstas por um modelo e os rótulos reais dos dados. A ideia por trás da função “sparse_categorical_crossentropy” é fornecer um sinal de erro que guia o processo de treinamento da rede neural, permitindo que os pesos sejam ajustados de forma a minimizar essa diferença entre as previsões do modelo e os rótulos reais. Ao minimizar a entropia cruzada (medida de dissimilaridade entre duas distribuições de probabilidade: a verdadeira distribuição dos dados e a distribuição predita por um modelo) durante o treinamento, o modelo é incentivado a produzir previsões mais precisas e confiáveis para as diferentes classes do problema de classificação (RAHMAN, 2023) .

3.2.3 Pré processamento

O pré-processamento de dados é fundamental no aprendizado de máquina, pois influencia diretamente a qualidade e o desempenho dos modelos. Isso se dá por varias razões importantes. Primeiramente, a limpeza dos dados é essencial, já que conjuntos reais frequentemente possuem ruídos, valores ausentes, *outliers* e inconsistências, o que pode prejudicar o desempenho dos modelos. O pré-processamento permite lidar com esses problemas, seja preenchendo valores faltantes, removendo *outliers* ou corrigindo erros nos dados (PADILHA; CARVALHO, 2017).

O *One-Hot Encoding* é uma técnica popular de pré-processamento de dados amplamente utilizada em tarefas de aprendizado de máquina e processamento de linguagem natural (ONE, 2024). Essa técnica é especialmente útil quando lidamos com variáveis categóricas, transformando-as em uma representação numérica adequada para algoritmos de aprendizado de máquina. No *One-Hot Encoding*, cada categoria única em uma variável categórica é representada por um vetor binário, onde apenas um dos elementos é 1 e os outros são 0, indicando a presença ou ausência da categoria. Como exemplo suponha que temos uma variável categórica chamada "Animal"com os valores: Gato, Cachorro, Pássaro. A Tabela 3.1 mostra o resultado apos o uso do *One-Hot Encoder* onde esses valores categóricos foram transformados em vetores binários, onde cada coluna corresponde a uma categoria.

Tabela 3.1 - Tabela ilustrativa do One-Hot Encoding.

Animal	**Gato**	**Cachorro**	**Pássaro**
Gato	1	0	0
Cachorro	0	1	0
Pássaro	0	0	1
Gato	1	0	0
Pássaro	0	0	1

Fonte: Da autora (2025).

Uma das vantagens do *One-Hot Encoding* é a capacidade de preservar a natureza discreta das variáveis categóricas sem atribuir-lhes uma ordem artificial. Isso significa que, ao converter categorias em vetores binários, não estamos impondo relações de ordem ou magnitude entre elas, o que é crucial em muitos contextos de aprendizado de máquina. Essa abordagem permite que algoritmos de aprendizado de máquina interpretem corretamente a natureza das categorias, sem assumir relações que não existem, evitando assim distorções nos resultados.

No entanto, o *One-Hot Encoding* também tem algumas limitações importantes a serem consideradas. Uma delas é o aumento na dimensionalidade dos dados após a transformação, especialmente em casos com muitas categorias únicas. Isso pode levar a problemas de eficiência computacional (ONE, 2024).

Além da limpeza e da transformação de dados, a redução da dimensionalidade é uma etapa importante no pré-processamento de dados para análise e modelagem em aprendizado de máquina. Esta etapa envolve o descarte de características que podem não ser relevantes ou que possam introduzir ruído nos modelos preditivos. A necessidade de reduzir a dimensionalidade surge especialmente em conjuntos de dados com muitas características, onde algumas delas

podem não contribuir significativamente para a capacidade preditiva do modelo.

A *feature importance* (importância das características) é uma técnica fundamental para entender quais atributos têm maior impacto nas previsões de um modelo. Essa abordagem atribui uma pontuação a cada característica com base na sua influência no desempenho geral do modelo. Essa pontuação pode ser calculada de várias maneiras, dependendo do algoritmo de aprendizado de máquina usado. Por exemplo, em modelos baseados em árvores de decisão, como o *Random Forest*, a importância das características é medida pelo ganho de informação ou pela redução na impureza das divisões ao longo das árvores (GERON, 2019).

A importância das características em uma rede neural não é diretamente visível como nos modelos baseados em árvores de decisão, mas existem métodos específicos para avaliar seus atributos, sendo a *Permutation Feature Importance* um exemplo relevante. Essa técnica é utilizada para medir a contribuição de cada variável em modelos de aprendizado de máquina, principalmente após o treinamento. O processo envolve embaralhar aleatoriamente os valores de uma variável enquanto mantém as demais constantes e, em seguida, observar o impacto na performance do modelo (como acurácia ou erro). Se a variável for relevante, a qualidade do modelo se deteriora significativamente; se for irrelevante, a performance permanece praticamente inalterada (ALTMANN et al., 2010).

Em síntese, o pré-processamento de dados garante que os dados estejam limpos, formatados corretamente e representados de maneira compreensível para os algoritmos de aprendizado de máquina. Um pré-processamento cuidadoso e adequado é essencial para modelos mais robustos, com melhor desempenho e generalização para novos dados (BATISTA, 2003).

3.2.4 Métricas de Avaliação

Métricas de desempenho são medidas cruciais para avaliar a eficácia de um modelo de aprendizado de máquina em uma determinada tarefa. Elas oferecem diferentes perspectivas sobre o quão bom é o desempenho do modelo e são fundamentais em diversos contextos de aplicação. Para compreendê-las melhor, é essencial entender os parâmetros que compõem essas métricas:

a) verdadeiros positivos (VP), correspondem aos casos em que o modelo previu corretamente a classe positiva ou relevante. Em resumo, são situações em que a classe real é positiva e o modelo acertou ao prever essa positividade;

b) verdadeiros negativos (VN**),** representam os casos em que o modelo previu

corretamente a classe negativa ou não relevante. Ou seja, são situações em que a classe real é negativa e o modelo acertou ao prever essa negatividade;

c) falsos positivos (FP), refletem os casos em que o modelo previu incorretamente a classe positiva. Isso ocorre quando o modelo indica que a classe é positiva, mas na verdade é negativa. São conhecidos como "erros tipo I";

d) falsos negativos (FN), são os casos em que o modelo previu incorretamente a classe negativa. Isso acontece quando o modelo indica que a classe é negativa, mas na verdade é positiva. São chamados de "erros tipo II".

Estes parâmetros são utilizados na avaliação de modelos de classificação em problemas de aprendizado supervisionado. Abaixo apresentamos as definições das métricas utilizadas nesse trabalho que fazem uso desses conceitos:

a) acurácia, é uma medida simples e direta da proporção de predições corretas em relação ao total de predições feitas pelo modelo, dada pela equação 3.1;

$$\text{acurácia} = \frac{\text{verdadeiros positivos} + \text{verdadeiros}}{\text{número total de predições}} \tag{3.1}$$

b) precisão, é a proporção de verdadeiros positivos (amostras corretamente classificadas como positivas) em relação ao total de predições positivas (incluindo falsos positivos). Ela mede a precisão das predições positivas do modelo. Neste trabalho, uma predição é considerada positiva quando o modelo prevê corretamente a cidade onde um indivíduo vai estudar. Isso significa que o modelo acerta ao identificar a cidade correta de destino para os estudantes. A equação 3.2 nos dá a equação para o cálculo da precisão;

$$\text{precisção} = \frac{\text{verdadeiros positivos}}{\text{verdadeiros positivos} + \text{falsos positivos}} \tag{3.2}$$

c) revocação, também conhecido como sensibilidade, é a proporção de verdadeiros positivos em relação ao total de amostras que são realmente positivas (verdadeiros positivos mais falsos negativos). Ele mede a capacidade do modelode encontrar todos os exemplos positivos.A equação 3.3, traz o cálculo da revocação;

$$\text{revocação} = \frac{\text{verdadeiros positivos}}{\text{verdadeiros positivos} + \text{falsos negativos}} \tag{3.3}$$

d) f1-Score é uma métrica que combina precisão e *revocação* em uma única medida. É a média harmônica dessas duas métricas e fornece um equilíbrio entre precisão e *revocação*. Com o F1-Score, temos o Macro F1-Score, que

calcula a média aritmética dos F1-Scores de todas as classes. Essa métrica é especialmente útil em cenários com desequilíbrio entre as classes, pois trata cada classe de forma igual, independentemente de sua frequência. A equação 3.4 mostra como o F1-Score é calculado.

$$\text{F1-Score} = \frac{2 \text{ x precisão x revocação}}{\text{precisão + revocação}} \tag{3.4}$$

Além dessas métricas, uma prática comum para avaliação robusta de modelos é o uso de validação cruzada. O *k-fold cross validation* é uma técnica popular que divide o conjunto de dados em k subconjuntos (ou *"folds"*). Em cada uma das k iterações, um subconjunto é utilizado como conjunto de validação, enquanto os k-1 subconjuntos restantes são usados como conjunto de treinamento. Esse processo é repetido k vezes, permitindo que cada subconjunto seja utilizado uma vez como conjunto de validação. As métricas de desempenho são então calculadas como a média das métricas obtidas em cada uma das k iterações, proporcionando uma avaliação mais precisa e reduzindo a variabilidade associada a uma única divisão aleatória dos dados.

Essas métricas são frequentemente utilizadas para avaliar e comparar diferentes modelos de aprendizado de máquina e identificar áreas de melhoria em um modelo específico. Dependendo do contexto e dos requisitos da aplicação, diferentes métricas podem ser mais relevantes (BISHOP, 2006).

3.3 Grafos

Um grafo é uma estrutura matemática formalmente definida como um par $G = (V, E)$, onde V é um conjunto de vértices (ou nós) e E é um conjunto de arestas, que representam conexões entre pares de vértices. Mais especificamente, uma aresta pode ser definida como um par ordenado (u, v) em um grafo direcionado ou um conjunto não ordenado $\{u, v\}$ em um grafo não direcionado, onde $u, v \in V$. Os vértices representam entidades, enquanto as arestas representam relações entre essas entidades. Abaixo listamos alguns dos tipos de grafos existentes, de acordo com DIESTEL (2016). A Figura 3.1 contém a ilustração de cada um dos tipos de grafos explicados abaixo:

a) grafos não direcionados, nesses grafos, as arestas não possuem uma direção específica, o que significa que a relação entre os vértices é simétrica. Em outras palavras, se há uma aresta que conecta o vértice A ao vértice B, essa mesma aresta implica uma conexão do vértice B ao vértice A. Esses grafos são úteis para modelar relações onde a

direção não é relevante, como em uma rua de mão dupla, onde os carros podem passar nas duas direções;

b) grafos direcionados,ao contrário dos grafos não direcionados, as arestas nesses grafos têm uma direção específica, indicando uma relação unidirecional entre os vértices conectados. No exemplo da rua, nesse tipo de grafo podemos imaginar uma via de mão única, onde os carros podem passar apenas para uma direção;

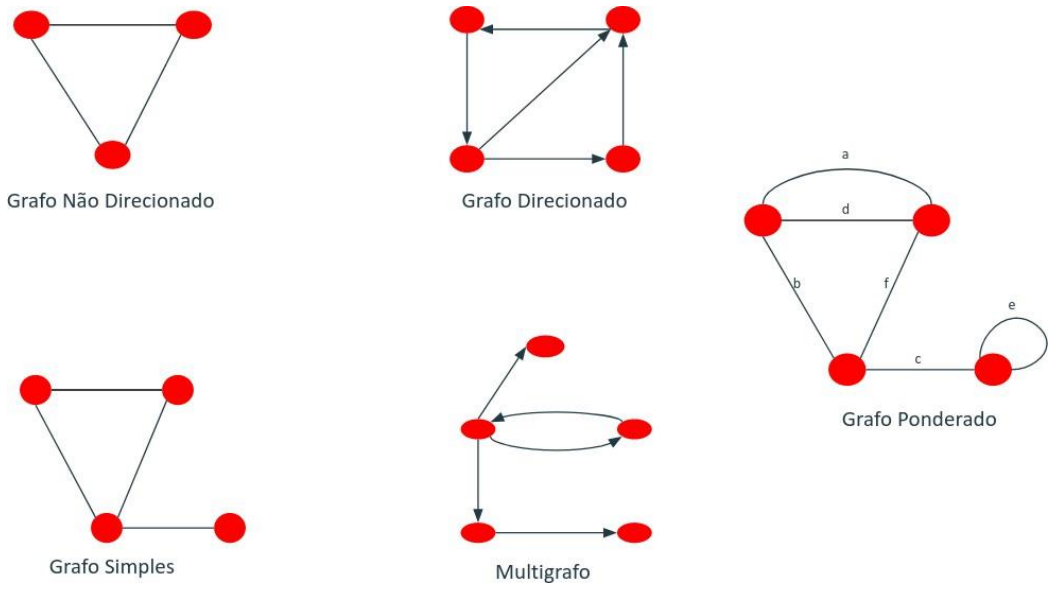
c) grafos simples, esses grafos não contêm laços nem arestas múltiplas entre os mesmos pares de vértices. Em outras palavras, cada par de vértices é conectado por no máximo uma única aresta. Isso simplifica a representação e análise do grafo, tornando-o adequado para muitas aplicações simples, como diagramas de redes sociais ou diagramas de amigos em redes sociais;

d) multigrafos, em contraste com os grafos simples, os multigrafos permitem múltiplas arestas entre os mesmos pares de vértices. Isto significa que é possível ter várias conexões entre os mesmos vértices, representando diferentes tipos ou intensidades de relações. Por exemplo, em uma rede de transporte público, onde podem existir várias linhas de ônibus entre duas estações, um multigrafo pode ser usado para representar essas relações complexas de forma mais precisa;

e) grafos ponderados, nestes grafos as arestas têm pesos associados, que podem representar diversas medidas quantitativas, como distâncias, custos, tempos de viagem, etc. Esses pesos adicionam uma dimensão adicional à estrutura do grafo, permitindo uma modelagem mais precisa de sistemas complexos. Por exemplo, em uma rede de estradas, os pesos das arestas podem representar as distâncias entre as interseções, enquanto em uma rede social, os pesos podem representar a força das conexões entre os usuários.

É importante notar que um grafo pode possuir múltiplas características simultaneamente. Por exemplo, um grafo pode ser tanto direcionado quanto ponderado, dependendo das propriedades específicas de suas arestas e vértices. Neste projeto, para o grafo de estudantes, utilizamos multigrafos ponderados e direcionados, nos quais cada vértice representa uma localização específica. As arestas, por sua vez, contêm informações sobre os estudantes e são direcionadas do local de nascimento para a localização onde eles escolheram estudar. Já no grafo de médicos, empregamos multigrafos ponderados e não direcionados, onde os vértices representam localidades e cada aresta corresponde a um médico que transitou entre essas localidades.

Figura 3.1 - Tipos de grafos.



Fonte: Da autora (2025).

3.4 Análise de Redes Sociais

A Análise de Redes Sociais (ARS) é uma abordagem para investigar estruturas sociais usando teoria de grafos (POWELL, 2015). Os nós podem representar as pessoas, e as arestas (ou links) podem representar as conexões entre elas, o grafo pode ser não direcionado ou direcionado dependendo do problema. ARS é um bom método para identificar padrões dentro de uma rede. Quanto maior a rede, mais relevante pode ser este tipo de análise (POWELL, 2015). Neste trabalho, utilizamos a centralidade de grau para analisar como a mobilidade dos estudantes de medicina e dos médicos evolui ao longo do tempo. Para isso, no caso dos estudantes, representamos os locais como vértices e as arestas correspondem à um estudante e sua migração cidade natal para o local onde escolheu estudar. No caso dos médicos, os locais também são representados como vértices, mas as arestas correspondem à um medico e a sua movimentação entre essas localidades.

O **grau de centralidade** é uma medida simples, ele calcula quantos vizinhos um nó possui em relação ao total de conexões possíveis. É calculado como o número de arestas que cada nó (v) possui. Se o grafo for direcionado, apenas as arestas que saem do nó (deg_{out}) são consideradas, também conhecido como grau externo (BORBA, 2013). A Equação 3.5 mostra a fórmula para grafos não direcionados, onde n é o número de nós no grafo e deg é o grau,

quantidade de arestas, do nó, enquanto a Equação 3.6 mostra as fórmulas para grafos direcionados. Para o grafo direcionado, é utilizado apenas o grau de saída (deg_{out}).

$$C_D(v_i) = \frac{\deg(v_i)}{n-1} \quad (3.5)$$

$$C_D(v_i) = \frac{\deg_{out}(v_i)}{n-1} \quad (3.6)$$

Desta forma, neste capítulo apresentamos os conceitos utilizados para o desenvolvimento deste trabalho. A seguir, esta detalhado a metodologia adotada para conduzir o estudo e alcançar os objetivos propostos.

4 METODOLOGIA

Este capítulo apresenta a metodologia utilizada para investigar o problema da mobilidade dos estudantes de medicina no Brasil. Após uma revisão sistemática para encontrar métodos e tecnologias a serem utilizados para enfrentar o problema apresentado, algoritmos de aprendizado de máquina foram testados e avaliados para obter uma boa precisão de onde os alunos tendem a ir com base em suas características. Em seguida, foi realizada uma análise de redes sociais para avaliar as medidas da rede, e então interpretar os resultados e avaliar se os objetivos da pesquisa foram cumpridos.

A seguir, apresentaremos os detalhes do desenvolvimento deste trabalho, começando pelos conjuntos de dados utilizados, os algoritmos de aprendizado de máquina e a análise de redes sociais.

4.1 Bases de dados

Para desenvolver as análises, contamos com dois conjuntos principais de dados. O primeiro conjunto de dados é a base de **Estudantes**, que contém os dados dos discentes de medicina registrados de 1995 até 2014. Estes dados são provenientes dos sistemas do INEP e do SUS e foram fornecidos pelos mesmos. O segundo conjunto de dados é a base de **Profissionais**, que contém as informações de vínculo de trabalho dos médicos e os respectivos estabelecimentos de saúde a cada mês dos anos 2012 até o ano 2021. A base de **Profissionais** é oriunda e foi fornecida pelo Cadastro Nacional de Estabelecimentos de Saúde (CNES). A base de dados dos **Estudantes** inclui dados como origem, destino e características do indivíduo e foi empregada tanto na predição quanto na análise de redes, enquanto a base de dados dos **Profissionais**, que inclui informações sobre os profissionais em diferentes instituições de saúde, refletindo sua mobilidade e atuação ao longo dos anos, foi exclusivamente utilizada na análise de redes. Essa abordagem permitiu uma visão integrada da formação e atuação dos profissionais de saúde no Brasil.

Os dados obtidos foram tratados seguindo princípios de pré-processamento de dados citados anteriormente. Para aprimorar a qualidade da base, foi realizada uma busca na base dos **Profissionais** para excluir indivíduos que não se enquadravam na categoria de profissionais de medicina, pois ela tinha dados de diferentes profissionais da área da saúde. Adicionalmente, foi introduzida uma coluna de data, permitindo a geração de fluxos de movimentação a partir dessa

informação. A base de **Profissionais** possui 1.516.317 entidades e 7 atributos e a de **Estudantes** 101.212 entidades e 14 atributos. Os Quadros 4.1 e 4.2 mostram os dados de cada uma das bases de dados **Estudantes** e **Profissionais**, respectivamente.

Quadro 4.1 - Base de Estudantes.

Nome	Tipo	Descrição
cpf	string	Registro de número individual (CPF)
idade	int	Idade do indivíduo
sexo	int	Sexo do indivíduo. Definido como 0 ou 1, onde 0 é masculino e 1 é feminino
raça	int	Cor ou raça declarada pelo indivíduo, pode ser definida como 0: Não declarado, 1: Branco, 2: Preto, 3: Marrom, 4: Amarelo, 5: Indígena, 6: Nenhuma informação disponível
co_uf_nasc	int	Código IBGE do estado onde o indivíduo foi registrado
co_uf_ies	int	Código IBGE do estado onde o indivíduo estudou medicina
idh_uf_nasc	float	Valor do índice de desenvolvimento humano do estado de nascimento do indivíduo
idh_uf_nasc	float	Valor do índice de desenvolvimento humano do estado onde o indivíduo estudou medicina
dist	int	Distancia entre o local de nascimento e onde o indivíduo estudou medicina
lat_nasc	float	Latitude do local onde o indivíduo nasceu
lon_nasc	float	Longitude do local onde o indivíduo nasceu
lat_ies	float	Latitude do local onde o indivíduo estudou medicina
lon_ies	float	Longitude do local onde o indivíduo estudou medicina
dt_inicio_curso	Date	Data de quando o indivíduo começou a estudar medicina

Fonte: Da autora (2025).

Quadro 4.2 - Base de Profissionais.

Nome	Tipo	Descrição
cpf	string	Registro de número individual (CPF)
cargo	string	Código e descrição da especialidade médica do indivíduo
competencia	string	Período em que o indivíduo ficou na localidade citada
lat	float	Latitude do local onde o indivíduo atende
lon	float	Longitude do local onde o indivíduo atende
datatime	Date	Data criada para geração de mapas
cod_uf	int	Código IBGE do estado onde o médico trabalha

Fonte: Da autora (2025).

O atributo *co_uf_ies* corresponde aos códigos dos estados, e foi usado como atributo classe para os modelos testados. Desta forma, queremos prever em qual Estado o indivíduo irá estudar a partir dos dados de sua origem. Na seção 4.2.1, apresentaremos mais sobre as informações dos dados antes e após o tratamento.

O gráfico da Figura 4.1 é um gráfico de distribuição de classes, onde o eixo X representa as diferentes classes presentes no conjunto de dados **Estudantes**, e no eixo Y o número de instâncias (ou ocorrências) pertencentes a cada uma dessas classes, as classes são referentes aos Estados da federação. Esse tipo de gráfico é útil para verificar se há uma distribuição equilibrada ou desequilibrada. Classes com um número similar de instâncias auxiliam a evitar que os algoritmos de aprendizado apresentem viés para uma classe específica. Por outro lado, se uma classe possui significativamente mais instâncias (classe majoritária) e outra muito menos (classe minoritária), pode ocorrer um problema de desbalanceamento, afetando o desempenho de modelos de aprendizado de máquina, principalmente em tarefas de classificação. A análise dessa distribuição é essencial para orientar a metodologia, ajudando a escolher estratégias adequadas, como técnicas de balanceamento ou o uso de métricas específicas, como o F1-score, que são menos impactadas pelo desbalanceamento (ZHANG; ZHOU, 2014).

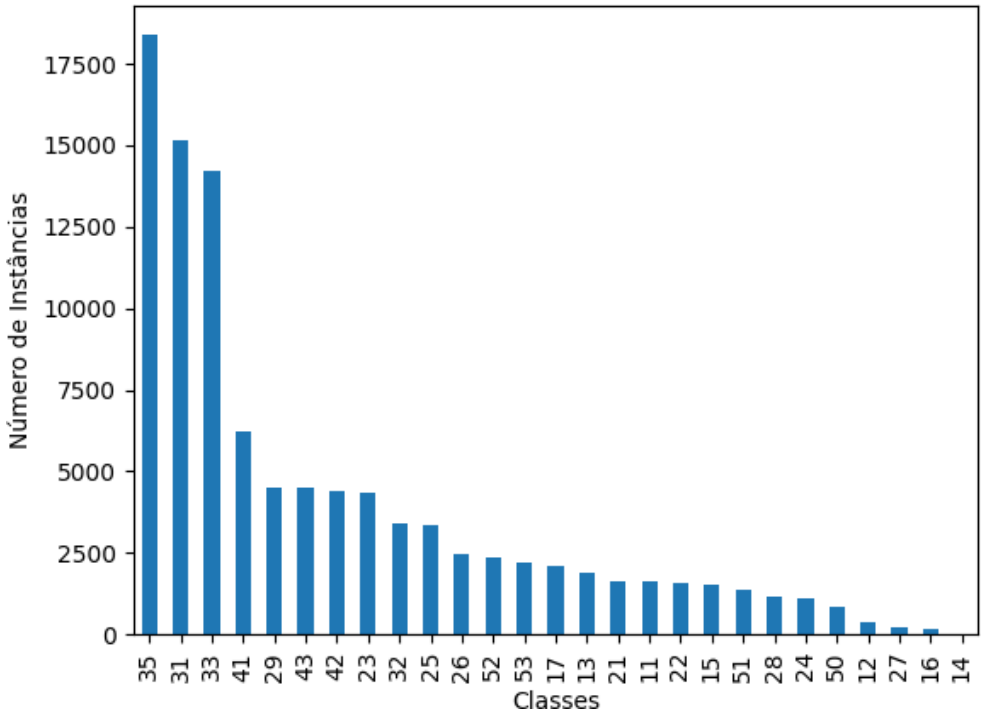
Podemos observar no gráfico na Figura 4.1 um alto grau de desbalanceamento. Algumas classes possuem um número significativamente maior de instâncias comparadas a outras. Este desbalanceamento pode ser observado, por exemplo, nas classes 35(MG) e 31(SP), que têm mais de 17.500 e 15.000 instâncias, respectivamente, enquanto as classes 16(AP) e 14(RR) possuem menos de 500 instâncias cada. Isto pode ser explicado pelo desequilíbrio na concentração de escolas de medicina em determinadas regiões. No apêndice A é apresentada uma relação dos estados e os códigos do Instituto Brasileiro de Geografia e Estatística (IBGE).

4.2 Estratégias de Tratamento, Modelagem e Análise dos Dados

Este trabalho usou algoritmos de aprendizado de máquina supervisionados, tendo em vista que o conjunto de dados contém o resultado esperado para cada entrada (rótulo), no caso deste trabalho, um estado onde o estudante irá estudar medicina. Mais especificamente, testamos os seguintes algoritmos: *Random Forest Classifier* e Rede Neural. Esses algoritmos foram escolhidos por possuírem estruturas diferentes e, consequentemente, oferecerem abordagens distintas para o problema. O *Random Forest Classifier* é um método baseado em árvores de decisão, o que o torna mais interpretável e robusto contra ruídos, além de lidar bem com dados desbalanceados. Já a Rede Neural, por sua arquitetura baseada em camadas de

neurônios interconectados, tem a capacidade de capturar padrões mais complexos e não lineares nos dados, podendo oferecer um melhor desempenho dependendo da distribuição das variáveis.

Figura 4.1 - Distribuição de classes (UFs)



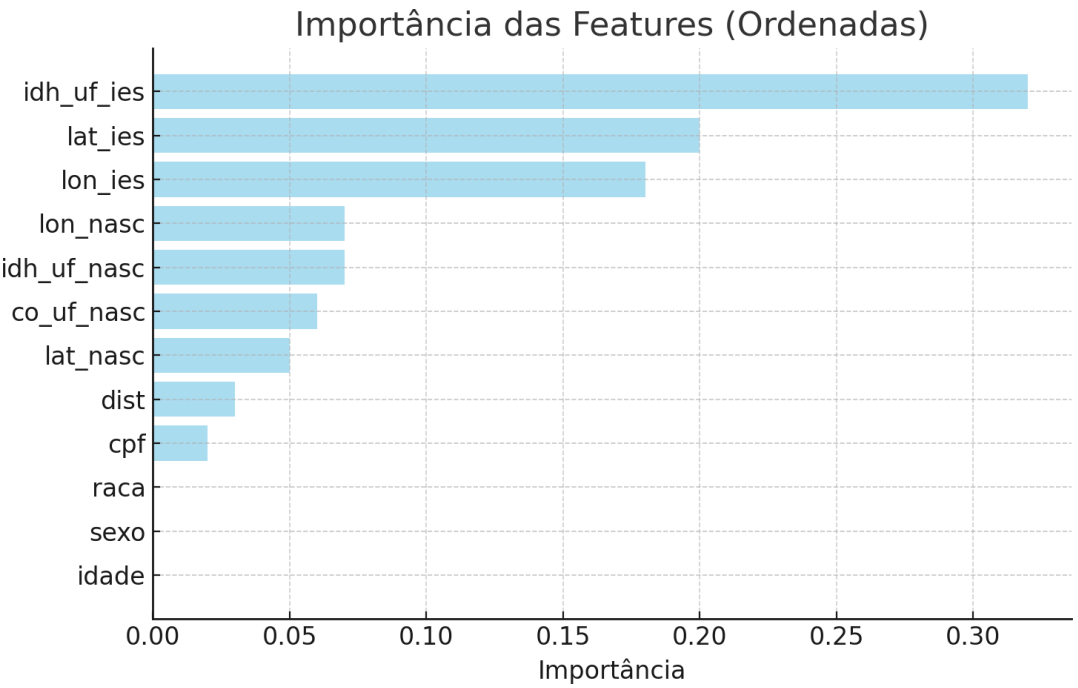
Fonte: Da autora (2025).

4.2.1 Tratamento dos Dados

Na etapa de pré-processamento, foram removidos os atributos *cp f* , *idade*, *sexo*, *raca* e *dt_inicio_curso* que não contribuíram significativamente para a predição do destino dos discentes de medicina. Essa decisão foi baseada na análise dos valores obtidos através das técnicas, *Feature Importance* (GERON, 2019) para o **Random Forest**, e da *Permutation Feature Importance* (ALTMANN et al., 2010) para a Rede Neural, que revelou a baixa relevância desses atributos. As Figuras 4.2 e 4.3 ilustram estes resultados.

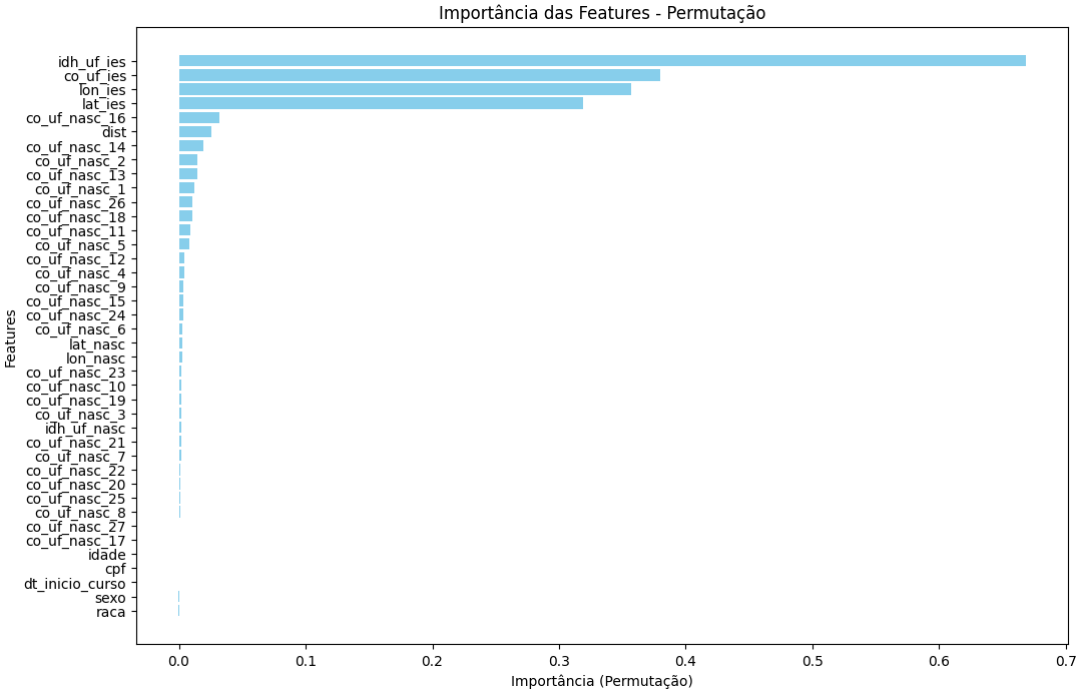
Decidimos manter apenas os dados de *idh_u f _ies*, *lat_ies* e *lon_ies*, devido à sua maior importância, conforme demonstrado no gráfico. Além disso, optamos por manter os dados de idh_u f _nasc, lat_nasc, lon_nasc e co_u f _nasc, pois esses dados correspondem aos dados de alta importância, mas do local de nascimento do estudante. Dessa forma, preservamos informações relevantes tanto do local de nascimento quanto do local de estudo, garantindo uma análise mais completa e precisa.

Figura 4.2 - Importância das classes pelo *Random Forest*



Fonte: Da autora (2025).

Figura 4.3 - Importância das classes pela Rede Neural

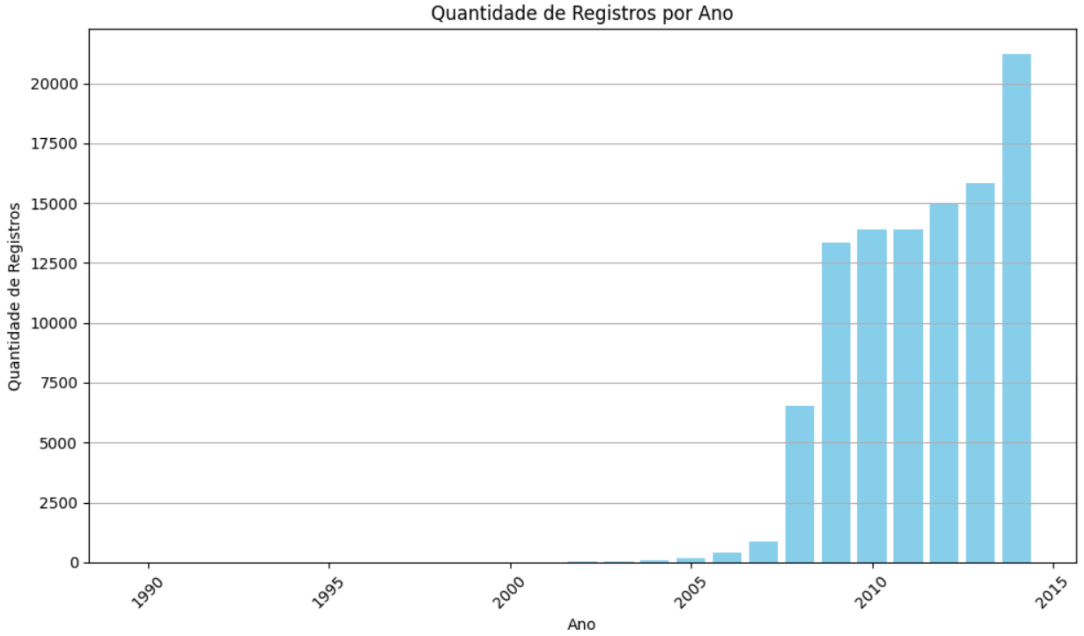


Fonte: Da autora (2025).

Durante essa etapa, identificamos um período com baixa quantidade de dados na base **Estudantes**. Essa informação está ilustrada na Figura 4.4. Com base nessa observação, para

garantir a consistência das análises de redes sociais, delimitamos o período analisado entre 2008 e 2014, quando há uma quantidade mais consistente de registros disponíveis.

Figura 4.4 - Quantidade registros por ano



Fonte: Da autora (2025).

.

A técnica de *one-hot encoding* foi aplicada ao atributo referente aos códigos dos Estados da IES onde o discente estuda/estudou e do estado de nascimento (*co_u f _ies* e *co_u f _nasc*) apenas para o modelo de rede neural. Esta técnica converte os códigos dos Estados (UFs) em atributos individuais, permitindo que o algoritmo não interprete os códigos das UFs como valores numéricos relacionados pela ordem de grandeza.

É importante ressaltar que os códigos das UFs são variáveis originalmente representadas por números inteiros. Embora numericamente expressos, esses códigos não seguem um padrão específico, o que pode levar modelos como as redes neurais a interpretarem incorretamente este atributo, relacionando localizações geográficas entre estados com códigos numericamente próximos. Esta codificação busca diminuir a possibilidade de interpretações inadequadas, garantindo uma representação apropriada das variáveis categóricas para a rede neural.

Por outro lado, a técnica de *one-hot encoding* não é necessária para o *Random Forest*, devido ao fato de ser um algoritmo baseado em árvores de decisão que é capaz de lidar com variáveis categóricas de forma natural, sem a necessidade de transformar os códigos em atributos individuais. As árvores de decisão tratam as variáveis categóricas diretamente e não fazem suposições sobre a ordem ou magnitude dos códigos das UFs, evitando interpretações

incorretas sem a necessidade de codificação.

4.2.2 Algoritmos para Modelos de Aprendizado de Máquina

A decisão de selecionar esses dois algoritmos é fundamentada em considerações estratégicas. O *Random Forest Classifier*, um algoritmo tradicional, destaca-se por sua robustez e capacidade de lidar eficientemente com uma variedade de conjuntos de dados. Por outro lado, redes neurais têm a capacidade intrínseca de aprender padrões complexos e representações abstratas nos dados. Essa complexidade é especialmente útil em contextos nos quais as relações entre variáveis são intrincadas e não lineares.

O *Random Forest*, com sua simplicidade e eficácia em muitos casos, serve como uma referência sólida. Enquanto isso, a Rede Neural, mais complexa, busca capturar nuances mais sutis nos dados. Essa abordagem dupla visou fornecer *insights* sobre a adequação dos modelos em diferentes cenários, possibilitando uma análise mais completa e informada das previsões geradas. A utilização desses dois algoritmos permitiu uma comparação, abordando diferentes características de modelos de aprendizado de máquina.

O conjunto de dados foi dividido em conjuntos de treinamento e teste usando o método *K-fold Cross Validation*, uma técnica de validação usada para avaliar a performance de um modelo de aprendizado de máquina, como explicado na seção 3.2.4

A construção do modelo de Rede Neural foi realizada utilizando o *TensorFlow*. A rede proposta é do tipo densa, composta por três camadas. A camada de entrada possui 35 neurônios, correspondendo ao número de características na base de dados **Estudantes**, após as transformações realizadas. A camada de saída contém 27 neurônios, representando o total de classes (estados) a serem previstas. Cada neurônio da camada de saída fornece a probabilidade da entrada pertencer a um estado específico, viabilizando a previsão em um problema de classificação multi-classe.

Entre a entrada e a saída, há uma camada oculta com 31 neurônios, definida como a média aritmética entre o número de neurônios das camadas de entrada e saída. Essa configuração busca equilibrar a capacidade da rede de capturar padrões, permitindo que a rede aprenda tanto representações intermediárias quanto interações complexas entre as variáveis. A escolha da arquitetura foi baseada nas orientações apresentadas em um documento que resume

boas práticas na definição de redes neurais[1].

Para a função de perda, foi utilizada *sparse_categorical_crossentropy*, adequada para problemas de classificação multi-classe com rótulos fornecidos como inteiros. O otimizador escolhido foi o *adam*, conhecido por sua eficiência computacional e baixa necessidade de ajuste de hiperparâmetros, resultando em uma convergência mais rápida e estável.

O modelo foi treinado por 10 (dez) épocas, um valor inicial que balanceia o tempo de treinamento e a performance, permitindo verificar a capacidade de aprendizado e convergência da rede. Dependendo dos resultados, ajustes podem ser feitos posteriormente para otimizar a acurácia e eficiência do modelo.

O *Random Forest* foi construído com um *max_depth* de 8 e *n_estimators* de 100. O modelo foi treinado com o conjunto de treinamento e avaliado no conjunto de testes.
Para o modelo de *Random Forest*, foi realizado um *Grid Search* para otimização dos hiperparâmetros. Esta abordagem foi adotada devido à natureza flexível e extensa da *Random Forest*, onde é crucial encontrar a combinação ideal de parâmetros para maximizar o desempenho do modelo. Já para a rede neural, optou-se por não utilizar o *grid search* devido à sua complexidade computacional e ao grande número de hiperparâmetros envolvidos. Em vez disso, a definição dos parâmetros foi baseada em documentações e aplicações encontradas durante a pesquisa (UFRJ, 2024).

4.2.3 Análise de Redes Sociais e Geoespacial

Com os dados tratados, foram gerados dois grafos a partir das bases **Estudantes** e **Profissionais**, cada um representando um conjunto específico de dados, onde os nós correspondem aos estados brasileiros. Para construir esses grafos, cada linha do *dataset* foi lida e processada, e os nós e arestas foram adicionados utilizando o pacote *NetworkX*[2] versão 2.8. No grafo da base **Estudantes**, as arestas são direcionadas, apontando para o estado de destino do curso e contêm informações dos estudantes que elas representam. No grafo da base **Profissionais**, as arestas são não direcionadas, refketindo a mobilidade dos médicos entre estados e também incluem dados dos profissionais. Esses grafos permitiram a comparação das variações nos índices de centralidade e das dinâmicas de mobilidade ao longo dos anos.

---

[1] http://www.nce.ufrj.br/labic/downloads/dicas_cfg_rna.pdf, acessado em 30/06/2024
[2] https://networkx.org/documentation/stable/index.html, acessado em 20/10/2023

A Análise de Redes Sociais (ARS) usando grafos é uma boa abordagem para entender a estrutura e a dinâmica das interações em uma rede social. Uma métrica importante nesse contexto é a centralidade de nós, que ajuda a identificar quais nós são mais importantes ou influentes na rede. A técnica de centralidade aplicada neste trabalho foi a centralidade de grau. Ela mede a importância de um nó pelo número de conexões diretas que possui. Nós com alta centralidade de grau são essenciais para a robustez da rede, pois sua remoção pode causar grandes impactos. Além disso, a centralidade de grau ajuda a entender por que alguns nós possuem mais conexões e o que pode ser feito para tornar essa distribuição mais igualitária. Para calcular esta centralidade foi utilizado novamente o pacote *NetworkX* na versão 2.8, muito aplicado em estudos de redes complexas.

A Análise Geoespacial compreende a criação de mapas de calor para definir as localizações geográficas de onde a maior parte dos estudantes de medicina se origina e para onde vão estudar.

4.2.4 Análise dos Resultados

Os resultados foram analisados em termos de métricas de desempenho: acurácia, precisão, revocação e F1-Score de cada modelo. O *k-fold cross validation* também foi utilizado na avaliação de desempenho, garantindo uma validação mais robusta e confiável dos modelos.

Foram incluídas visualizações gráficas que exploram a relação entre variáveis específicas no conjunto de dados, como o IDH nos estados de nascimento e estudo.

A metodologia proposta visou direcionar a comparação do desempenho de diferentes modelos de aprendizado de máquina na tarefa de predição de classes, bem como reportar resultados de análises de redes sociais a partir dos grafos gerados pela mobilidade dos estudantes de medicina.

5 RESULTADOS

Neste capítulo, iremos apresentar os resultados finais obtidos através da aplicação de técnicas de Aprendizado de Máquina e Análise de Redes Sociais.

5.1 Rede Neural Artificial e *Random Forest*

Para conduzir esta análise, utilizamos a base de dados denominada **Estudantes**, que contém informações sobre alunos de medicina. Essa base foi dividida em um conjunto de treino e outro de teste, conforme descrito na metodologia, para avaliar a capacidade do modelo em prever o local onde os indivíduos cursarão medicina. A variável alvo, ou seja, aquela que estamos tentando prever, é representada pela classe, que indica o Estado escolhido pelos alunos para estudo. A Tabela 5.1 apresenta os resultados das médias das classes após as execuções dos modelos e a avaliação com os dados de teste.

Tabela 5.1 - Relatório de Classificação dos Modelos de Aprendizado de Máquina

Modelo	**Acurácia**	**Precisão**	**Recall**	**F1-Score**
Rede Neural	97,90%	98,23%	97,90%	97,98%
Random Forest	98,36%	98,78%	98,36%	98,32%

Fonte: Da autora (2025).

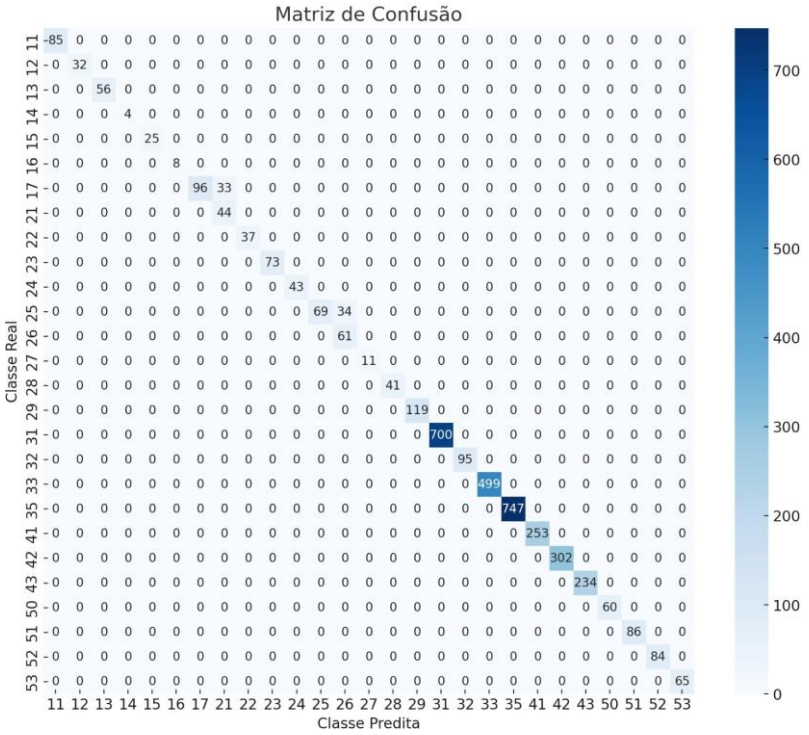
Ambos os modelos apresentaram resultados robustos na tarefa de classificação dos dados da base **Estudantes**. A Rede Neural Artificial obteve uma acurácia média de 97,90% e um Macro F1-Score médio de 97,98%, demonstrando alta precisão e equilíbrio entre precisão e *recall*. Por outro lado, o *Random Forest* alcançou uma acurácia média de 98,36% e um Macro F1-Score médio de 98,32%, indicando bons resultados em termos de precisão e capacidade de generalização.

As Figuras 5.1 e 5.2 mostram a matriz de confusão de cada um dos modelos.

Dado o objetivo de prever o destino dos estudantes de medicina com alta precisão e considerando os resultados obtidos, o modelo *Random Forest* se destaca como a escolha mais eficaz. Com uma acurácia superior e desempenho consistente em todas as métricas avaliadas, o *Random Forest* demonstrou ser a técnica mais adequada para esta tarefa específica de classificação. Uma melhor acurácia e F1-Score são relevantes porque indicam que o modelo é capaz de realizar classificações precisas e equilibradas, fundamentais para decisões informadas

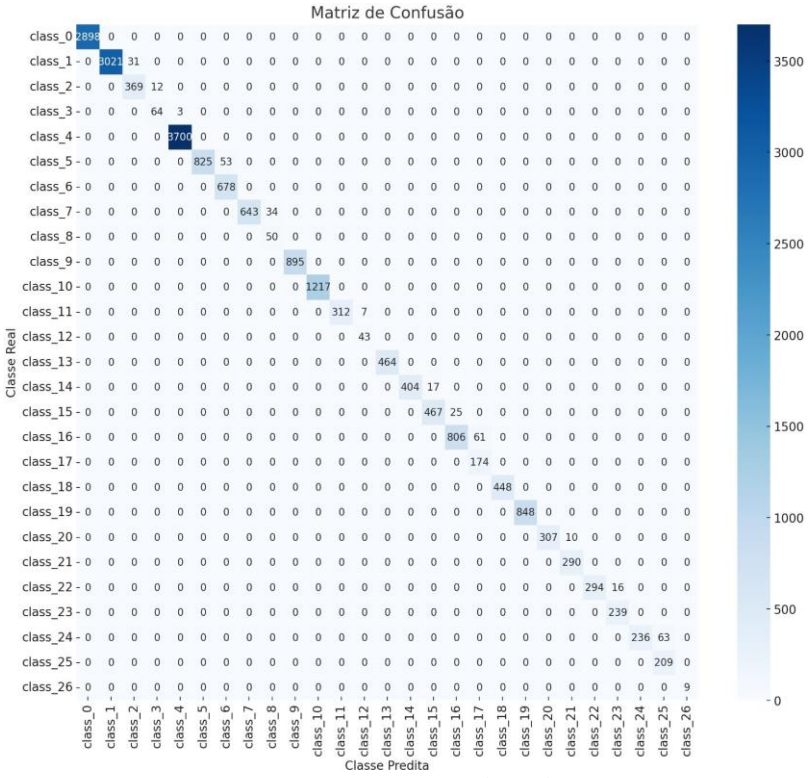
e eficazes.

Figura 5.1 - Matriz Confusão Random Forest



Fonte: Da autora (2025).

Figura 5.2 - Matriz Confusão Rede Neural



Fonte: Da autora (2025).

Estes resultados consolidam a eficácia tanto da Rede Neural Artificial quanto do *Random Forest* na tarefa de classificação dos dados da base **Estudantes**. Essas métricas são essenciais para orientar a escolha do modelo mais adequado, considerando a importância relativa de acurácia e eficiência para a previsão dos destinos dos estudantes de medicina.

5.2 Análise de Redes Sociais

Para avaliar as métricas de rede geradas pelo deslocamento de estudantes de medicina, utilizamos a técnica de avaliação de centralidade de grau, abrangendo o grau total, de saída e de entrada. O grau total representa o número total de conexões incidentes em um nó específico, o grau de saída é o número de conexões que partem desse nó para outros, e o grau de entrada é o número de conexões que apontam para esse nó.

Na base de dados dos **Estudantes**, analisamos os três tipos de centralidade de grau a partir do grafo multigrafo, direcionado e ponderado gerado. Já na base dos **Profissionais**, foi realizada apenas a análise do grau total, pois o grafo é um multigrafo não direcionado e ponderado, o que impede o cálculo das outras métricas de centralidade.

O período considerado para análise na base de **Estudantes** abrange os anos de 2008 a 2014, uma vez que em períodos anteriores havia uma quantidade limitada de dados disponíveis, o que poderia comprometer a análise. Já o período considerado para análise na base de **Profissionais** abrange os anos de 2012 a 2024, pois a base compreendia apenas este período.

Os Anexos B e C contém os dados usados para gerar os gráficos que serão apresentados.

5.2.1 Base Estudantes

A seguir, são descritos os resultados das análises de centralidades de graus dos estudantes. Esta análise foi realizada utilizando o grafo gerado a partir da base **Estudantes**.

Para avaliar melhor os resultados da centralidade desses dados utilizamos como referência os números de faculdades de medicina e sua distribuição pelo Brasil no ano de 2012. Estes dados foram encontrados em um relatório que faz parte de um projeto Programa de Apoio Institucional ao Desenvolvimento do Sistema Único de Saúde (PROADI-SUS), do Ministério da Saúde. Esse critério permite contextualizar a mobilidade dos estudantes, avaliando se a concentração de instituições em determinadas regiões influencia a centralidade.

A Tabela 5.2 apresenta estes dados.

Tabela 5.2 - Dados sobre Cursos de Medicina no Brasil.

Estado	**População (Censo 2010)**	**Cursos de Medicina**	**Total de Vagas**	**Habitantes/Vaga**
Acre	732.793	1	80	9.160
Alagoas	3.120.922	2	150	20.806
Amapá	668.689	1	60	11.145
Amazonas	3.480.937	3	340	10.238
Bahia	14.021.432	8	693	20.233
Ceará	8.448.055	7	692	12.208
Distrito Federal	2.562.963	4	306	8.376
Espírito Santo	3.512.672	5	500	7.025
Goiás	6.004.045	4	330	18.194
Maranhão	6.569.683	3	214	30.699
Mato Grosso	3.033.991	3	182	16.670
Mato Grosso do Sul	2.449.341	3	240	10.206
Minas Gerais	19.595.309	30	2.564	7.642
Pará	7.588.078	4	390	19.457
Paraíba	3.766.834	7	620	6.076
Paraná	10.439.601	12	1.007	10.367
Pernambuco	8.796.032	6	650	13.532
Piauí	3.119.015	4	342	9.120
Rio de Janeiro	15.993.583	18	2.147	7.449
Rio Grande do Norte	3.168.133	3	236	13.424
Rio Grande do Sul	10.695.532	11	952	11.235
Rondônia	1.560.501	4	230	6.785
Roraima	451.227	1	80	5.640
Santa Catarina	6.249.682	10	514	12.159
São Paulo	41.252.160	36	3.081	13.389
Sergipe	2.068.031	3	210	9.848
Tocantins	1.383.453	4	302	4.581
Brasil	**190.755.799**	**197**	**17.112**	**11.147**

Fonte: Estudantes de Medicina e Médicos no Brasil: Números Atuais e Projeções(2013).

5.2.1.1 Grau de Centralidade de Entrada

O grau de centralidade de entrada em uma rede nos fornece informações sobre a quantidade de conexões direcionadas que entram em um nó específico na rede. Em outras palavras, ele indica quantas arestas estão apontando para o nó em questão.

A centralidade de entrada mede o número de conexões direcionadas para um nó específico, indicando o quão atraente ou receptivo um estado é como destino para estudantes de medicina vindos de outras regiões. Aplicada à rede de mobilidade de estudantes, essa medida revela os principais polos formativos e os estados que mais atraem alunos de fora, independente de sua origem.

A Tabela 5.3 apresenta os resultados da centralidade de entrada para cada Estado entre os anos de 2008 e 2014. Ao analisar o ano de 2012 na Tabela 5.2, observa-se que a centralidade de entrada dos estados brasileiros, combinada com a oferta de cursos de medicina, podemos ter um panorama detalhado sobre a mobilidade dos estudantes na área da saúde.

O Estado de Minas Gerais alcançou um alto valor para a centralidade de entrada, de 4.268, o que indica uma alta capacidade de atrair estudantes de medicina. Este estado também possui uma oferta robusta de 30 cursos de medicina e 2.564 vagas disponíveis. Isso demonstra que, além de ser um ponto central na rede de mobilidade, Minas Gerais oferece um número significativo de oportunidades para os futuros médicos.

O Estado de São Paulo apresentou com uma centralidade de entrada de 5.292, se posicionando como o principal polo de atração, refletindo sua grande quantidade de cursos (36) e vagas (3.081). Esses dados evidenciam que São Paulo não apenas contém uma grande quantidade de instituições, mas também desempenha um papel crucial na formação de médicos no Brasil, tornando-se a principal escolha para estudantes.

Outros Estados com altas centralidades de entrada incluem o Rio de Janeiro, com 3.992, e o Espírito Santo, com 1.039. O Rio de Janeiro oferece 18 cursos e 2.147 vagas, mostrando-se o Rio de Janeiro como o 3º estado mais atraente para estudantes de medicina.

Por outro lado, estados como Roraima, com uma centralidade de entrada de apenas 1 e 80 vagas, refletem uma baixa atratividade para estudantes. Isso pode ser atribuído à sua oferta limitada de cursos, que pode não atender à demanda local ou regional. Acre e Amapá, ambos com centralidade de entrada de 1 e 80 e 60 vagas respectivamente, compartilham a mesma limitação, indicando um cenário onde os estudantes tendem a buscar oportunidades em estados com mais cursos disponíveis.

Em resumo, os dados de centralidade de entrada, quando comparados com a oferta de cursos de medicina, revelam que estados com maior centralidade atraem significativamente mais estudantes, evidenciando a importância de uma infraestrutura educacional robusta para fomentar a formação médica no Brasil. Essa análise destaca a relevância de Minas Gerais e São Paulo como centros de atração e formação na área da saúde, enquanto estados com menor centralidade enfrentam desafios em termos de captação de estudantes. Esta desigualdade regional acentua a "fuga de cérebros", especialmente em estados com baixa centralidade de entrada, onde muitos estudantes não retornam após a graduação. Mas, para isto, precisamos analisar a retenção dos estudantes nos locais onde se graduaram, o que foge do escopo deste trabalho.

Tabela 5.3 - Dados Centralidade Entrada por Estado de 2008 a 2014.

Estado	2008	2009	2010	2011	2012	2013	2014
MG	4268.0	4268.0	4268.0	4268.0	4268.0	4268.0	4268.0
RJ	3992.0	3992.0	3992.0	3992.0	3992.0	3992.0	3992.0
ES	1039.0	1039.0	1039.0	1039.0	1039.0	1039.0	1039.0
SP	5292.0	5292.0	5292.0	5292.0	5292.0	5292.0	5292.0
TO	133.0	166.0	111.0	111.0	111.0	111.0	111.0
RO	182.0	228.0	152.0	152.0	152.0	152.0	152.0
AC	26.0	32.0	22.0	22.0	22.0	22.0	22.0
RR	2.0	0.0	1.0	1.0	1.0	1.0	1.0
PA	249.0	312.0	216.0	216.0	216.0	216.0	216.0
AM	286.0	345.0	240.0	240.0	240.0	240.0	240.0
AP	0.0	0.0	22.0	22.0	22.0	22.0	22.0
MA	210.0	185.0	185.0	185.0	185.0	185.0	185.0
PE	331.0	294.0	294.0	294.0	294.0	294.0	294.0
PB	419.0	397.0	397.0	397.0	397.0	397.0	397.0
CE	590.0	517.0	517.0	517.0	517.0	517.0	517.0
SE	148.0	134.0	134.0	134.0	134.0	134.0	134.0
RN	147.0	129.0	129.0	129.0	129.0	129.0	129.0
PI	213.0	187.0	187.0	187.0	187.0	187.0	187.0
BA	600.0	525.0	525.0	525.0	525.0	525.0	525.0
AL	0.0	25.0	25.0	25.0	25.0	25.0	25.0
PR	2423.0	2423.0	2423.0	2423.0	2423.0	2423.0	2423.0
SC	1897.0	1897.0	1897.0	1897.0	1897.0	1897.0	1897.0
RS	2117.0	2117.0	2117.0	2117.0	2117.0	2117.0	2117.0
MS	166.0	166.0	166.0	166.0	166.0	166.0	166.0
DF	553.0	553.0	553.0	553.0	553.0	553.0	553.0
GO	689.0	689.0	689.0	689.0	689.0	689.0	689.0
MT	328.0	328.0	328.0	328.0	328.0	328.0	328.0

Fonte: Da autora (2025).

5.2.1.2 Grau de Centralidade de Saída

A centralidade de saída é uma medida que indica a quantidade de estudantes que deixam um estado para se deslocar para outros locais, seja para estudar ou para trabalhar. Valores mais altos de centralidade de saída sugerem um fluxo intenso de estudantes para fora do estado, enquanto valores mais baixos podem indicar uma retenção maior de estudantes ou uma oferta limitada de oportunidades fora do estado.

A Tabela 5.4 apresenta os resultados da centralidade de saída dos estudantes dos estados brasileiros. Quando esses dados são analisados em conjunto com a distribuição dos cursos de medicina em 2012, mostrada na Tabela 5.2, é possível obter uma perspectiva interessante sobre como os estudantes se deslocam e se distribuem pelo Brasil.

Minas Gerais e São Paulo também aparecem como os estados com a maior centralidade de saída. Minas Gerais apresenta uma centralidade de saída constante de 4.955, enquanto São Paulo tem 5.743. Esses estados não apenas atraem estudantes, mas também têm um fluxo significativo de estudantes que se deslocam para outras regiões, indicando uma forte mobilidade dos estudantes formados nesses locais. O Rio de Janeiro e o Espírito Santo também mostram valores elevados de centralidade de saída, com 2.838 e 1.054, respectivamente. Isso sugere que, apesar de serem pontos de atração, muitos estudantes também deixam esses estados para estudar em outros locais. Por outro lado, estados como Tocantins, Roraima e Acre têm valores baixos de centralidade de saída (103, 11 e 36, respectivamente), indicando uma menor mobilidade dos estudantes para fora desses estados.

Analisando a relação com a oferta de cursos de medicina em 2012, vemos Minas Gerais com 30 cursos e 2.564 vagas, mostrando-se como um estado relevante na formação médica e também na mobilidade, tanto como destino quanto como origem de estudantes. São Paulo, com 36 cursos e 3.081 vagas, tem a maior centralidade de saída e também a maior capacidade de formar e enviar médicos para outros estados. O Rio de Janeiro e o Espírito Santo também têm uma boa oferta de cursos (18 e 5, respectivamente), o que pode contribuir para a saída de estudantes que buscam oportunidades em outros locais.

Por outro lado, estados com menor número de cursos, como Roraima (1 curso, 80 vagas) e Acre (1 curso, 80 vagas), têm baixa centralidade de saída, apesar de ser maior do que a centralidade de entrada. Isto sugere que os estudantes tendem a buscar educação médica em locais onde a oferta é maior e não necessariamente mais próximo.

Em resumo, a análise da centralidade de saída, em combinação com a oferta de

cursos de medicina, revela que estados com maior centralidade de saída também possuem uma infraestrutura educacional robusta, permitindo a formação de médicos que buscam oportunidades em outros locais. Estados como Minas Gerais e São Paulo não só atraem estudantes, mas também formam médicos que contribuem para a mobilidade nacional.

Tabela 5.4 - Dados Centralidade Saída por Estado de 2008 a 2014.

Estado	2008	2009	2010	2011	2012	2013	2014
MG	4955.0	4955.0	4955.0	4955.0	4955.0	4955.0	4955.0
RJ	2838.0	2838.0	2838.0	2838.0	2838.0	2838.0	2838.0
ES	1054.0	1054.0	1054.0	1054.0	1054.0	1054.0	1054.0
SP	5743.0	5743.0	5743.0	5743.0	5743.0	5743.0	5743.0
TO	103.0	129.0	86.0	86.0	86.0	86.0	86.0
RO	175.0	219.0	146.0	146.0	146.0	146.0	146.0
AC	36.0	45.0	30.0	30.0	30.0	30.0	30.0
RR	11.0	0.0	9.0	9.0	9.0	9.0	9.0
PA	298.0	373.0	253.0	253.0	253.0	253.0	253.0
AM	255.0	317.0	212.0	212.0	212.0	212.0	212.0
AP	0.0	0.0	27.0	27.0	27.0	27.0	27.0
MA	211.0	185.0	185.0	185.0	185.0	185.0	185.0
PE	380.0	336.0	336.0	336.0	336.0	336.0	336.0
PB	231.0	202.0	202.0	202.0	202.0	202.0	202.0
CE	688.0	603.0	603.0	603.0	603.0	603.0	603.0
SE	133.0	120.0	120.0	120.0	120.0	120.0	120.0
RN	153.0	134.0	134.0	134.0	134.0	134.0	134.0
PI	202.0	177.0	177.0	177.0	177.0	177.0	177.0
BA	660.0	581.0	581.0	581.0	581.0	581.0	581.0
AL	0.0	55.0	55.0	55.0	55.0	55.0	55.0
PR	2697.0	2697.0	2697.0	2697.0	2697.0	2697.0	2697.0
SC	1435.0	1435.0	1435.0	1435.0	1435.0	1435.0	1435.0
RS	2306.0	2306.0	2306.0	2306.0	2306.0	2306.0	2306.0
MS	156.0	156.0	156.0	156.0	156.0	156.0	156.0
DF	376.0	376.0	376.0	376.0	376.0	376.0	376.0
GO	913.0	913.0	913.0	913.0	913.0	913.0	913.0
MT	292.0	292.0	292.0	292.0	292.0	292.0	292.0

Fonte: Da autora (2025).

5.2.1.3 Grau de Centralidade Total

O grau de centralidade total em um grafo representa a importância de um nó (neste caso, um estado) em termos de conexões com outros nós. Valores mais altos de grau de centralidade total indicam que o estado é um ponto focal em uma rede, possuindo muitos vínculos, o que sugere uma significativa influência e fluxo de estudantes.

Os dados da centralidade de grau total dos estados brasileiros, apresentados na Tabela 5.5 variam significativamente. São Paulo, com uma centralidade de grau total de 11.035, se destaca como o estado com a maior centralidade. Isso é corroborado pelo fato de ser o estado com o maior número de cursos de medicina (36) e vagas (3.081), refletindo uma capacidade formativa. A alta centralidade sugere que muitos estudantes se deslocam para São Paulo para buscar formação e, após a conclusão dos cursos, muitos podem optar por se deslocar para outras regiões, mantendo o estado como um centro educacional e profissional.

Tabela 5.5 - Dados Centralidade Total por Estado de 2008 a 2014.

Estado	2008	2009	2010	2011	2012	2013	2014
MG	9223.0	9223.0	9223.0	9223.0	9223.0	9223.0	9223.0
RJ	6830.0	6830.0	6830.0	6830.0	6830.0	6830.0	6830.0
ES	2094.0	2094.0	2094.0	2094.0	2094.0	2094.0	2094.0
SP	11035.0	11035.0	11035.0	11035.0	11035.0	11035.0	11035.0
TO	237.0	295.0	198.0	198.0	198.0	198.0	198.0
RO	358.0	447.0	299.0	299.0	299.0	299.0	299.0
AC	62.0	78.0	52.0	52.0	52.0	52.0	52.0
RR	13.0	0.0	11.0	11.0	11.0	11.0	11.0
PA	548.0	685.0	469.0	469.0	469.0	469.0	469.0
AM	541.0	662.0	453.0	453.0	453.0	453.0	453.0
AP	0.0	0.0	49.0	49.0	49.0	49.0	49.0
MA	422.0	370.0	370.0	370.0	370.0	370.0	370.0
PE	711.0	630.0	630.0	630.0	630.0	630.0	630.0
PB	650.0	600.0	600.0	600.0	600.0	600.0	600.0
CE	1279.0	1120.0	1120.0	1120.0	1120.0	1120.0	1120.0
SE	281.0	255.0	255.0	255.0	255.0	255.0	255.0
RN	300.0	264.0	264.0	264.0	264.0	264.0	264.0
PI	416.0	364.0	364.0	364.0	364.0	364.0	364.0
BA	1260.0	1107.0	1107.0	1107.0	1107.0	1107.0	1107.0
AL	0.0	81.0	81.0	81.0	81.0	81.0	81.0
PR	5120.0	5120.0	5120.0	5120.0	5120.0	5120.0	5120.0
SC	3332.0	3332.0	3332.0	3332.0	3332.0	3332.0	3332.0
RS	4423.0	4423.0	4423.0	4423.0	4423.0	4423.0	4423.0
MS	323.0	323.0	323.0	323.0	323.0	323.0	323.0
DF	930.0	930.0	930.0	930.0	930.0	930.0	930.0
GO	1603.0	1603.0	1603.0	1603.0	1603.0	1603.0	1603.0
MT	621.0	621.0	621.0	621.0	621.0	621.0	621.0

Fonte: Da autora (2025).

Minas Gerais (MG), tem centralidade de 9.223. Possui uma quantidade significativa de cursos (30) e vagas (2.564). A centralidade elevada indica que MG não apenas forma um grande número de estudantes, mas também desempenha um papel crucial na mobilidade, funcionando como um ponto de origem para profissionais de saúde. O Rio de Janeiro e o

Espírito Santo apresentam centralidades de 6.830 e 2.094, respectivamente. Ambos os estados têm uma influência significativa na formação médica e no fluxo de estudantes. O RJ, em particular, é um centro importante, com 18 cursos e 2.147 vagas, enquanto o ES, com 5 cursos e 500 vagas, mantém uma centralidade que indica um papel considerável na formação de profissionais que podem migrar para outros estados.

Por outro lado, estados como Roraima (RR) e Acre (AC) apresentam centralidades bem mais baixas (13 e 62, respectivamente). A baixa centralidade sugere uma menor participação na rede de formação médica, provavelmente devido a uma oferta limitada de cursos e vagas. Esses estados também têm populações menores, o que contribui para a escassez de médicos.

A análise da centralidade de grau total deve ser feita em conjunto com a quantidade de cursos de medicina e a relação habitantes/vaga em cada estado,como notamos na Tabela 5.2. Estados como Minas Gerais e São Paulo apresentam uma população grande em relação ao número de vagas de medicina, resultando em uma taxa de habitantes/vaga mais baixa (7.642 e 13.389, respectivamente). Isto indica que, embora sejam grandes formadores de médicos, a alta demanda por educação médica é atendida com um número proporcional de vagas. Em contrapartida, estados como Maranhão (30.699 habitantes/vaga) e Alagoas (20.806 habitantes/vaga) têm uma centralidade baixa e uma relação habitantes/vaga elevada, indicando que a capacidade de formar médicos é limitada em comparação à demanda populacional. Isso pode resultar em uma migração de estudantes para estados com maior oferta de cursos, como SP e MG.

A centralidade de grau total tem implicações diretas para a formação e a distribuição de profissionais de saúde. A alta centralidade em estados como SP e MG pode indicar uma concentração de recursos e oportunidades em detrimento de estados com menor centralidade. Isso levanta questões sobre a equidade na formação médica e na distribuição de profissionais. Para melhorar a distribuição de médicos pelo Brasil, seria prudente considerar políticas que incentivem a abertura de cursos de medicina em estados com baixa centralidade, onde a oferta é escassa, e onde a população pode se beneficiar de um maior acesso à saúde.

Estas mudanças e as estabilidades nas centralidades a partir de 2008 podem ser atribuídas a resultados de políticas públicas implementadas pelos governos. Essas políticas incluem investimentos em educação para melhorar a qualidade do ensino em regiões específicas, programas de desenvolvimento regional destinados a promover o crescimento

econômico e social em áreas menos desenvolvidas, bem como incentivos educacionais, como bolsas de estudo e financiamento estudantil. Além disso, a expansão da infraestrutura educacional, como a abertura de novas universidades ou campus universitários em determinadas regiões, pode ter aumentado o acesso à educação superior localmente, reduzindo a necessidade de migração de estudantes para outras áreas em busca de oportunidades educacionais. Essas políticas podem ter impactado as preferências dos estudantes e, consequentemente, influenciado as tendências de migração de estudantes entre as regiões, refletindo-se nas mudanças e estabilidades observadas nas centralidades de entrada e saída nos dados (OLIVEIRA et al., 2019).

Em conclusão, a centralidade de grau total dos estados na rede de mobilidade de estudantes de medicina revela um panorama complexo da formação médica no Brasil. Enquanto estados como São Paulo e Minas Gerais se destacam pela sua capacidade de formar profissionais, a análise também evidencia as disparidades na oferta de cursos e na distribuição de médicos entre as diferentes regiões. Essa informação é crucial para a formulação de políticas públicas que busquem equilibrar a oferta de educação médica e melhorar a distribuição de profissionais de saúde em todo o país.

5.2.2 Base de Dados de Profissionais

A seguir, são apresentados os resultados da análise de centralidade de grau dos profissionais. Para esta análise, foi gerado um grafo a partir da base **Profissionais**. Esta base contém os dados de médicos em diferentes instituições de saúde, com os locais e períodos que atuou em cada um deles.

5.2.3 Grau de Centralidade

A análise da centralidade total na rede de profissionais revela o quão conectados e influentes os estados são no contexto da mobilidade e distribuição desses profissionais ao longo dos anos. Nessa abordagem, a centralidade total, mede a quantidade de conexões de um estado com os demais, destacando aqueles que concentram mais profissionais ou que funcionam como polos de atração e movimentação no sistema de saúde. A Tabela 5.6 apresenta os valores da conectividade dos profissionais.

Os resultados indicam que São Paulo (SP), com uma centralidade de 148.656 em todos os anos analisados, é o estado com maior relevância. Esta alta centralidade sugere que SP é um ponto central na rede de profissionais de saúde no Brasil, atuando não apenas como local de origem e destino, mas também como ponto de concentração desses profissionais. A grande oferta de infraestrutura hospitalar, oportunidades de carreira e especialização podem justificar essa predominância. A concentração de profissionais em SP pode, no entanto, indicar um problema de desigualdade na distribuição, afetando o acesso à saúde em regiões menos centrais.

Minas Gerais (MG), com centralidade de 89.761, e Rio de Janeiro (RJ), com 50.971, também desempenham papéis importantes na rede. Esses estados são grandes centros formadores e empregadores de profissionais da saúde, reforçando seu papel como regiões estratégicas na mobilidade de médicos.

Outros estados da região sul, como Rio Grande do Sul (RS), Santa Catarina (SC) e Paraná (PR), apresentam centralidades altas (80.573, 53.197 e 66.157, respectivamente). Isso indica que o Sul do Brasil é uma região relevante na retenção e circulação de profissionais, com uma boa distribuição interna de recursos e oportunidades.

Por outro lado, estados como Acre (AC), Amapá (AP) e Roraima (RR) apresentam centralidades baixas (2.293, 1.055 e 1.764, respectivamente). Esses valores sugerem uma menor integração dessas unidades na rede de profissionais, o que pode refletir uma dificuldade em atrair e manter profissionais de saúde. A baixa centralidade nesses estados é preocupante, pois pode estar associada à uma falta de infraestrutura e oportunidades de carreira, além de uma menor oferta de serviços de saúde.

Estados do Norte e Nordeste, como Pará (PA), Maranhão (MA) e Piauí (PI), apresentam centralidades intermediárias (12.380, 12.249 e 6.464, respectivamente). Isso sugere que, embora não sejam polos tão relevantes quanto São Paulo ou Minas Gerais, ainda desempenham um papel importante na rede, especialmente como pontos de retenção de profissionais formados na própria região. No entanto, a mobilidade para outras regiões mais centrais pode representar uma dificuldade para manter esses profissionais em seus estados de origem.

Distrito Federal (DF), com uma centralidade de 5.646, desempenha um papel peculiar na rede. Embora seja um centro administrativo, sua centralidade é relativamente baixa comparada a estados como SP e RJ. Isso pode indicar que o fluxo de profissionais é mais concentrado em

áreas metropolitanas com maior densidade populacional e infraestrutura médica.

A comparação entre a centralidade total e a infraestrutura médica dos estados revela desigualdades importantes. Estados com alta centralidade, como SP e MG, concentram grande parte dos profissionais, enquanto estados com baixa centralidade, como AP e RR, enfrentam desafios significativos para garantir a oferta de serviços de saúde. Essa desigualdade pode impactar o acesso da população a cuidados médicos adequados, especialmente em regiões mais afastadas dos grandes centros.

Em suma, a análise da centralidade total da base de profissionais destaca uma concentração de médicos e profissionais de saúde nas regiões sudeste e sul, especialmente em estados como São Paulo, Minas Gerais e Rio de Janeiro. Embora esses estados ofereçam grandes oportunidades, a concentração pode resultar em um desequilíbrio na oferta de serviços de saúde, prejudicando regiões periféricas e menos integradas na rede. Esses resultados sugerem a necessidade de políticas públicas que incentivem a distribuição mais equilibrada de profissionais de saúde, promovendo a retenção de médicos em estados com baixa centralidade e ampliando o acesso à saúde em todo o país.

Tabela 5.6 - Dados por Estado de 2012 a 2020

(Continua)

Estado	2012	2013	2014	2015	2016	2017	2018	2019	2020
SP	148.656	148.656	148.656	148.656	148.656	148.656	148.656	148.656	148.656
RJ	50.971	50.971	50.971	50.971	50.971	50.971	50.971	50.971	50.971
ES	26.965	26.965	26.965	26.965	26.965	26.965	26.965	26.965	26.965
MG	89.761	89.761	89.761	89.761	89.761	89.761	89.761	89.761	89.761
SC	53.197	53.197	53.197	53.197	53.197	53.197	53.197	53.197	53.197
RS	80.573	80.573	80.573	80.573	80.573	80.573	80.573	80.573	80.573
PR	66.157	66.157	66.157	66.157	66.157	66.157	66.157	66.157	66.157
AC	2.293	2.293	2.293	2.293	2.293	2.293	2.293	2.293	2.293
PA	12.380	12.380	12.380	12.380	12.380	12.380	12.380	12.380	12.380
AM	3.475	3.475	3.475	3.475	3.475	3.475	3.475	3.475	3.475
TO	3.232	3.232	3.232	3.232	3.232	3.232	3.232	3.232	3.232
RO	6.917	6.917	6.917	6.917	6.917	6.917	6.917	6.917	6.917
AP	1.055	1.055	1.055	1.055	1.055	1.055	1.055	1.055	1.055
RR	1.764	1.764	1.764	1.764	1.764	1.764	1.764	1.764	1.764
MA	12.249	12.249	12.249	12.249	12.249	12.249	12.249	12.249	12.249
BA	39.939	39.939	39.939	39.939	39.939	39.939	39.939	39.939	39.939
CE	38.191	38.191	38.191	38.191	38.191	38.191	38.191	38.191	38.191
SE	5.968	5.968	5.968	5.968	5.968	5.968	5.968	5.968	5.968
AL	5.023	5.023	5.023	5.023	5.023	5.023	5.023	5.023	5.023
PI	6.464	6.464	6.464	6.464	6.464	6.464	6.464	6.464	6.464
PE	34.102	34.102	34.102	34.102	34.102	34.102	34.102	34.102	34.102

Tabela 5.6 - Dados por Estado de 2012 a 2020.

(Conclusão)

Estado	2012	2013	2014	2015	2016	2017	2018	2019	2020
PB	14.457	14.457	14.457	14.457	14.457	14.457	14.457	14.457	14.457
RN	8.204	8.204	8.204	8.204	8.204	8.204	8.204	8.204	8.204
DF	5.646	5.646	5.646	5.646	5.646	5.646	5.646	5.646	5.646
MT	6.288	6.288	6.288	6.288	6.288	6.288	6.288	6.288	6.288
MS	6.975	6.975	6.975	6.975	6.975	6.975	6.975	6.975	6.975
GO	41.163	41.163	41.163	41.163	41.163	41.163	41.163	41.163	41.163

Fonte: Da autora (2025).

5.3 Mapas de Calor

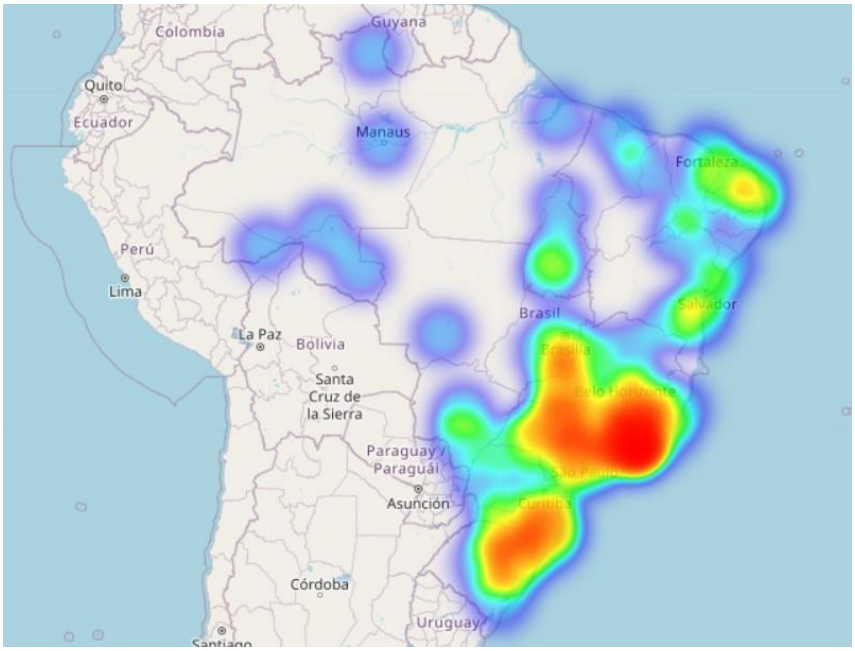
Os mapas de calor são ferramentas visuais utilizadas para representar a densidade de um fenômeno específico em uma área geográfica. Eles utilizam gradientes de cores para indicar concentrações, com cores quentes como vermelho e amarelo indicando altas densidades e cores frias como azul e verde indicando baixas densidades. Esta forma de visualização facilita a identificação de padrões e tendências espaciais, tornando-se especialmente útil em diversas áreas de estudo, incluindo a análise de dados educacionais (ANDRIENKO; ANDRIENKO, 2005).

Os mapas de calor apresentados a seguir (Figuras 5.3 e 5.4) foram gerados com os dados da base **Estudantes** e ilustram a distribuição de estudantes na América do Sul, com foco particular no Brasil, nos anos de 2008 e 2014. Através da comparação desses dois mapas, podemos observar como a densidade de estudantes evoluiu ao longo destes seis anos, permitindo uma análise detalhada das mudanças na distribuição geográfica dos locais de estudo. Este estudo busca entender melhor como o acesso e a concentração de instituições de ensino superior mudaram ao longo do tempo, refletindo os impactos de políticas educacionais e o desenvolvimento socioeconômico regional.

A análise da distribuição de estudantes ao longo dos anos 2008 e 2014 revela mudanças no cenário educacional do Brasil. Utilizando os mapas de calor criados Figuras 5.3 e 5.4, é possível observar a evolução na densidade de estudantes em diferentes regiões.

Na Figura 5.3, nota-se que em 2008, as regiões com maior densidade de estudantes estavam concentradas principalmente no Sudeste do Brasil. Esta área, destacada pelas cores vermelha e amarela, indica uma alta concentração de estudantes, refletindo a presença de reno madas instituições de ensino superior, como UNICAMP, USP e UFMG. No Nordeste do Brasil

Figura 5.3 - Mapa de calor 2008



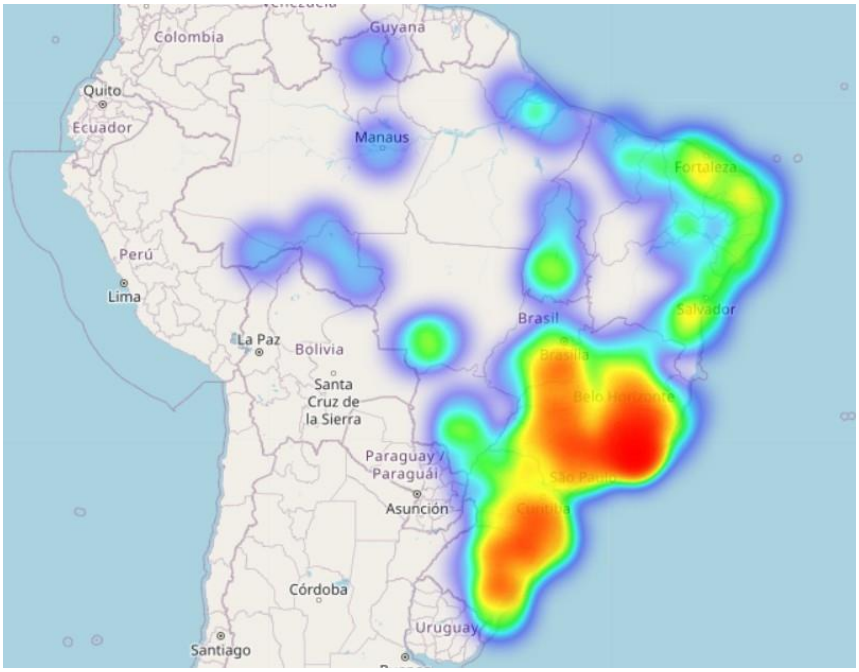
Fonte: Da autora (2025).

há uma presença moderada de estudantes, indicada pelas cores verde e azul claro. Outras partes do Brasil, incluindo o Norte e o Centro-Oeste, apresentavam menor densidade de estudantes, assim como os países vizinhos, como Argentina, Bolívia e Colômbia, que mostravam pouca ou nenhuma presença significativa de estudantes.

Já na Figura 5.4, referente ao ano de 2014, observa-se uma alteração na distribuição de estudantes. O Sudeste do Brasil continua sendo a região com maior concentração, com uma expansão das áreas em vermelho. Além disso, o Nordeste do Brasil mostra um aumento na densidade de estudantes em comparação com 2008, com mais áreas em verde e azul. O Centro-Oeste do Brasil também registra um aumento na presença de estudantes, embora em menor escala. Algumas áreas do Norte do Brasil apresentam uma densidade ligeiramente maior em comparação a 2008, mas ainda são relativamente baixas

Ao analisar esses mapas de calor destaca-se uma expansão do acesso ao ensino superior no Brasil entre 2008 e 2014, que demonstram uma evolução positiva no cenário educacional brasileiro, com um aumento na densidade de estudantes em várias regiões do país nesse período. Isso indica uma maior democratização do acesso ao ensino superior, refletindo o impacto de políticas públicas e o crescimento das instituições educacionais.

Figura 5.4 - Mapa de calor 2014



Fonte: Da autora (2025).

.

5.4 Distribuição Geográfica

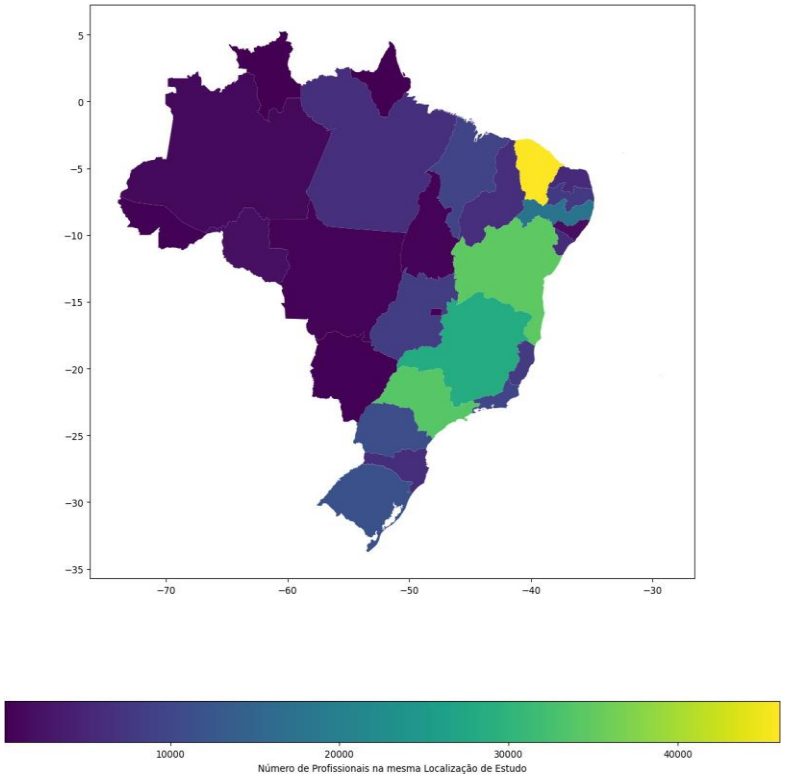
A Figura 5.5, gerada através dos dados da base **Profissionais**, mostra um mapa coroplético[2] do Brasil, representando a distribuição de profissionais que permaneceram na mesma localização de estudo para trabalhar. A escala de cores na parte inferior da imagem varia do roxo (indicando um número menor de profissionais) ao amarelo (indicando um número maior de profissionais). Essa escala facilita a identificação visual dos estados com mais profissionais que permaneceram no mesmo estado onde estudaram.

Ao analisar a Figura 5.5, podemos notar que os estados da região Norte (Acre, Amazonas, Rondônia, Roraima, Pará, Amapá, Tocantins) e Centro-Oeste (Mato Grosso, Mato Grosso do Sul, Distrito Federal) são representados principalmente por cores mais escuras (roxo e azul), indicando um número relativamente baixo de profissionais que permaneceram no estado onde estudaram.

---

[2] Mapa temático utilizado para representar variáveis quantitativas ou qualitativas por meio de diferentes cores ou tons aplicados a áreas geográficas. Cada região é preenchida com uma cor que reflete a intensidade, quantidade ou categoria de uma variável associada àquela área.

Figura 5.5- Distribuição de profissionais que permaneceram na mesma localização de estudo para trabalhar



Fonte: Da autora (2025).

Já a região Nordeste (Maranhão, Piauí, Ceará, Rio Grande do Norte, Paraíba, Pernambuco, Alagoas, Sergipe, Bahia) apresenta uma variação nas cores. Estados como Maranhão e Piauí estão em cores mais escuras (indicando menor número de profissionais), enquanto estados como Bahia, Ceará e Pernambuco estão em cores mais claras (verde e amarelo), sugerindo um número maior de profissionais que permaneceram no estado.

A região Sudeste (São Paulo, Rio de Janeiro, Minas Gerais, Espírito Santo) apresenta uma variação maior. São Paulo, por exemplo, está representado em uma cor verde, indicando um número substancial de profissionais que permaneceram no estado. Outros estados, como Minas Gerais, estão em cores intermediárias.

Por fim, a região Sul (Paraná, Santa Catarina, Rio Grande do Sul) apresenta cores mais escuras e intermediárias, indicando um número moderado de profissionais que permaneceram no mesmo estado de estudo.

O que essa análise nos mostra é que profissionais de estados com IDHs mais baixos podem estar migrando para estados com IDHs mais altos em busca de melhores oportunidades e condições de trabalho. A análise visual do mapa coroplético, combinada com os dados de

IDH, oferece uma visão de como a retenção de profissionais varia entre os estados do Brasil. Estados com IDHs mais altos tendem a reter mais profissionais, enquanto estados com IDHs mais baixos enfrentam desafios em manter seus profissionais locais. Uma análise mais aprofundada deve considerar outros fatores, como políticas estaduais, qualidade das instituições de ensino e condições econômicas de cada estado, para entender ainda mais a fundo essa escolha.

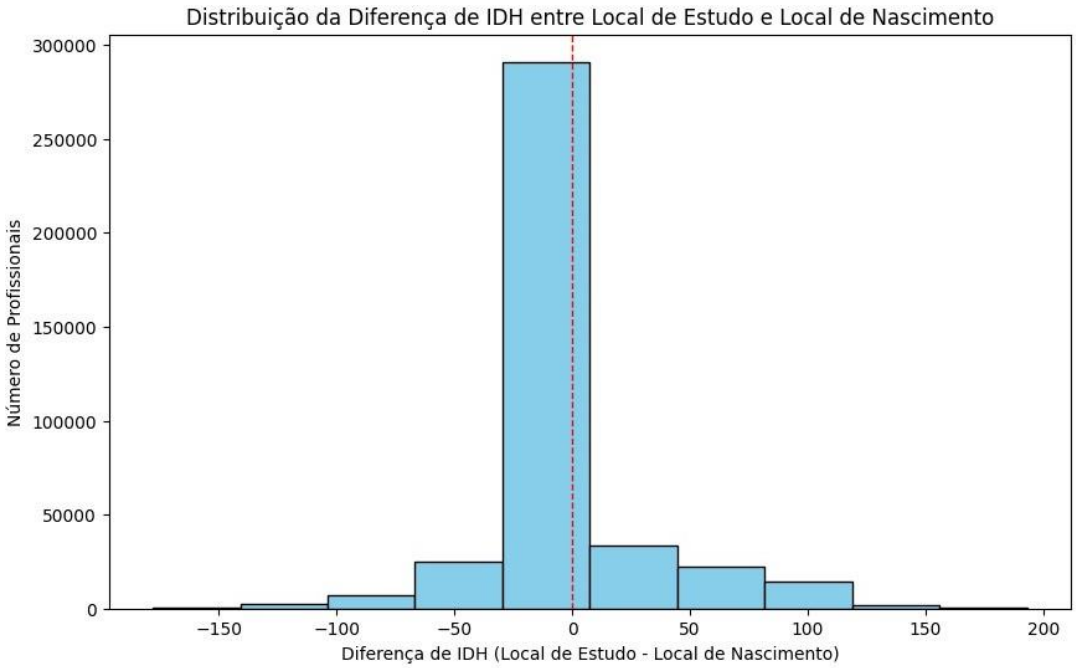
Tabela 5.7 - IDH dos Estados Brasileiros em 2021

Localidade	**IDH 2021**
Maranhão	676
Amapá	688
Pará	690
Piauí	690
Bahia	691
Roraima	699
Paraíba	698
Rondônia	700
Amazonas	700
Sergipe	702
Alagoas	684
Pernambuco	719
Acre	710
Rio Grande do Norte	728
Tocantins	731
Ceará	734
Mato Grosso	736
Goiás	737
Mato Grosso do Sul	742
Paraná	769
Espírito Santo	771
Rio Grande do Sul	771
Rio de Janeiro	762
Santa Catarina	792
Minas Gerais	774
São Paulo	806
Distrito Federal	814

Fonte: Da autora (2025).

Além disso, a Figura 5.6 mostra um histograma da diferença de IDH entre o local de trabalho e o local de nascimento para profissionais que permaneceram no mesmo local de estudo. A maioria dos profissionais trabalha no mesmo estado em que estudaram, resultando em uma diferença de IDH igual a zero. No entanto, há alguns casos em que a diferença de IDH é positiva ou negativa.

Figura 5.6- Distribuição da Diferença de IDH entre Local de Trabalho e Local de Nascimento para Profissionais no Mesmo Local de Estudo



Fonte: Da autora (2025).

6 DISCUSSÃO E CONCLUSÕES FINAIS

Este capítulo apresenta uma conclusão dos resultados obtidos, discutindo suas implicações e limitações. Além disso, destacam-se os impactos da centralidade das regiões no fluxo de estudantes e médicos. Por fim, são sintetizadas as principais contribuições do estudo, bem como suas restrições e possibilidades para pesquisas futuras.

6.1 Discussão

Nesta seção, são analisadas as constatações principais obtidas a partir dos resultados das análises de centralidade, dos mapas de calor, dos modelos de aprendizado de máquina e da análise de IDH relacionada à mobilidade profissional na área médica.

As análises de centralidade da base de **Estudantes** revelaram que a centralidade de entrada dos estados têm padrões na atração de estudantes de medicina. Alguns estados funcionam como polos, recebendo estudantes de diferentes origens e consolidando sua posição no sistema educacional. A centralidade geral no grafo gerado a partir da base de **Profissionais**, por ser não-direcionada, destacou a relevância de certas regiões como locais estratégicos na mobilidade dos médicos, sugerindo uma possível concentração de recursos e oportunidades de trabalho.

Os mapas de calor indicam uma expansão significativa do acesso ao ensino superior no Brasil entre 2008 e 2014. A maior densidade de estudantes em várias regiões do país sugere uma democratização do acesso ao ensino superior, impulsionada por políticas públicas e pelo crescimento das instituições educacionais (OLIVEIRA et al., 2019). Esta expansão reflete um cenário positivo no desenvolvimento educacional, o que pode influenciar diretamente a distribuição dos futuros profissionais no mercado de trabalho.

A análise do IDH entre o local de trabalho e o local de nascimento para profissionais que permaneceram no mesmo local de estudo, mostrou que há mais de 200.000 profissionais trabalham no mesmo local em que estudaram, resultando em uma diferença de IDH igual a zero. Existem casos de migração interna, com diferenças de IDH positivas ou negativas, sugerindo que alguns profissionais mudaram para outros estados. A média dos dados é centrada em zero, com poucas variações extremas, indicando que o IDH do estado natal tende a ser similar ao do estado onde o profissional escolhe trabalhar.

Em relação aos modelos de aprendizado de máquina, foram alcançados os seguintes

resultados: a Rede Neural Artificial obteve uma acurácia média de 97,90% e um Macro F1-Score médio de 97,98%, demonstrando precisão e equilíbrio entre precisão e recall. O *Random Forest* apresentou uma acurácia média de 98,36% e um Macro F1-Score médio de 98,32%, indicando bons resultados tanto em termos de precisão quanto de capacidade de generalização. Ambos os modelos se mostraram eficientes na previsão da mobilidade, com uma leve vantagem para o *Random Forest*. A diferença nos resultados sugere que o *Random Forest* pode ser mais robusto em relação a outliers e dados heterogêneos, enquanto a Rede Neural tem o potencial de capturar padrões mais complexos, desde que bem ajustada.

Por fim, as análises indicam que, embora exista uma correlação positiva entre o local de estudo e o local de trabalho, esse alinhamento não é absoluto. Políticas de incentivo local, como programas de fixação de médicos, podem influenciar significativamente a decisão dos profissionais e devem ser consideradas em análises futuras.

6.2 Conclusão final

Este trabalho teve como objetivo investigar a mobilidade de médicos no Brasil, aplicando análises de centralidade e modelos de aprendizado de máquina para compreender padrões de fluxo e identificar possíveis desajustes entre a formação acadêmica e o mercado de trabalho. A análise integrada desses dois sistemas permitiu uma avaliação abrangente sobre como e por que os profissionais da saúde migram entre regiões, buscando não apenas explicar as escolhas individuais, mas também fornecer subsídios para políticas públicas mais eficientes.

Primeiramente, foi conduzida uma análise exploratória da centralidade para medir a atratividade e a retenção de diferentes estados tanto para estudantes quanto para médicos. A centralidade de entrada dos estados evidenciou que alguns atuam como polos, recebendo estudantes de diversas origens, consolidando sua posição no sistema educacional e, por vezes, também na alocação profissional. A aplicação de modelos de aprendizado de máquina acrescentou uma dimensão preditiva importante à pesquisa. Os resultados mostraram que tanto a Rede Neural Artificial quanto o *Random Forest* foram eficazes na previsão da mobilidade dos médicos. Esses modelos, embora diferentes em suas abordagens, confirmaram sua utilidade para prever padrões de mobilidade, com o *Random Forest* se mostrando mais generalizável e a Rede Neural mais adequada para capturar padrões complexos quando bem ajustada.

Apesar dos avanços apresentados, o trabalho tem algumas limitações. Primeiro, nem todos os fatores relevantes para a mobilidade foram capturados, como a qualidade de vida e políticas públicas específicas de cada estado, que poderiam afetar as decisões profissionais. Além disso, as Redes Neurais demandam alto poder computacional e um processo intensivo de ajuste, o que pode limitar sua aplicação prática em cenários com menos recursos. Por fim, como o estudo se concentra no contexto brasileiro, a generalização dos resultados para outros países deve ser feita com cautela, considerando as diferenças socioeconômicas e culturais.

Para dar continuidade a esta pesquisa, alguns trabalhos futuros são sugeridos. Primeiramente, seria interessante expandir o conjunto de dados para incluir variáveis qualitativas, como entrevistas com médicos, a fim de compreender melhor as motivações subjetivas por trás da escolha de um local de trabalho. Além disso, a aplicação de algoritmos adicionais, como *Graph Neural Networks (GNNs)*, pode melhorar a precisão das previsões e explorar de maneira mais profunda a estrutura dos grafos. Uma análise das centralidades de intermediação e proximidade também pode ser útil para identificar estados que atuam como pontos estratégicos na mobilidade dos profissionais, oferecendo uma nova perspectiva sobre a eficiência da rede de formação e trabalho. Por fim, seria relevante analisar a evolução temporal da mobilidade para investigar se mudanças nas políticas públicas ao longo dos anos têm impacto nos padrões detectados.

Em resumo, este trabalho buscou contribuir para a literatura sobre mobilidade profissional e oferecer *insights* relevantes para a formulação de políticas públicas mais alinhadas entre a formação educacional e as demandas do mercado de trabalho. Para isso, utilizou técnicas da Ciência da Computação, como análise de grafos e aprendizado de máquina, para identificar padrões de mobilidade e potenciais desajustes entre a formação e a alocação profissional. Acredita-se que os resultados e reflexões apresentados possam inspirar novas pesquisas e orientar gestores públicos na implementação de estratégias mais eficientes para a retenção de profissionais em regiões prioritárias, promovendo uma melhor distribuição de recursos humanos na saúde e, consequentemente, um atendimento mais equitativo à população.

REFERÊNCIAS

ALTMANN, A. et al. Permutation importance: a corrected feature importance measure. **Bioinformatics**, v. 26, n. 10, p. 1340–1347, 04 2010. ISSN 1367-4803. Available from Internet: <https://doi.org/10.1093/bioinformatics/btq134>.

ANAND, S. et al. China's human resources for health: quantity, quality, and distribution. **The Lancet Elsevier**, v. 372, n. 9651, p. 1774–1781, out. 2008. Available from Internet: <https://doi.org/10.1016/S0140-6736(08)61363-X>.

ANDRIENKO, N.; ANDRIENKO, G. **Exploratory Analysis of Spatial and Temporal Data: A Systematic Approach**. Berlin, Heidelberg: Springer-Verlag, 2005. ISBN 3540259945.

ASGARI, F.; GAUTHIER, V.; BECKER, M. A survey on human mobility and its applications. 07 2013.

BATISTA, G. E. de A. P. A. **Pré-processamento de Dados em Aprendizado de Máquina Supervisionado**. Thesis (Doctorate) — ICMC-USP, 2003. Available from Internet: <https://edisciplinas.usp.br/pluginfile.php/4052836/mod_resource/content/4/mineracaodadosbiologicos-parte3.pdf>.

BERGEVIN1, Y. et al. Towards the triple aim of better health, better care and better value for canadians: transforming regions into high performing health systems. **Canadian Foundation for Healthcare Improvement**, 2015. Available from Internet: <https://www.mcgill.ca/familymed/files/familymed/regionalization_report_en_1.pdf>.

BISHOP, C. M. **Pattern Recognition and Machine Learning**. [S.l.]:Springer, 2006.

BODO, L. Aprendizagem de máquina para análise de indicadores em processos de software. 2015. Available from Internet: <https://repositorio.unesp.br/bitstream/handle/11449/154707/000869005.pdf?sequence=1>.

BORBA, E. M. **Medidas de Centralidade em Grafos e Aplicações em redes de dados**. 2013. 77 p. Available from Internet: <https://www.lume.ufrgs.br/bitstream/handle/10183/86094/000909891.pdf>.

CAMPOS, C. V. de A.; MALIK, A. M. Satisfação no trabalho e rotatividade dos médicos do programa de saúde da família. **Revista de Administração Pública**, v. 42, n. 2, p. 347–368, 2008. Available from Internet: <https://www.scielo.br/j/rap/a/kKH6BLCbVfMXrMk8vHLzT9S/?format=pdf&lang=pt>.

CELBIs¸, M. G. Applications of machine learning models in regional and demographic economic analysis: A literature survey. In: _____. **Labor Markets, Migration, and Mobility: Essays in Honor of Jacques Poot**. Singapore: Springer Singapore, 2021. p. 219–229. ISBN 978-981-15-9275-1. Available from Internet: <https://doi.org/10.1007/978-981-15-9275-1_ 10>.

COSTIGLIOLA, V. Mobility of medical doctors in cross-border healthcare. **EPMA Journal**, v. 2, n. 4, p. 333–339, 2011. Available from Internet: <https://www.ncbi.nlm.nih.gov/pmc/articles/PMC3405407/>.

DIESTEL, R. **Graph Theory**. 5. ed. [S.l.]: Springer, 2016. v. 173. ISBN 978-3-662-53621-6.

FENG, J. et al. Deepmove: Predicting human mobility with attentional recurrent networks. In: **Proceedings of the 2018 World Wide Web Conference**. Republic and Canton of Geneva, CHE: International World Wide Web Conferences Steering Committee, 2018. (WWW '18), p. 1459–1468. ISBN 9781450356398. Available from Internet: <https://doi.org/10.1145/3178876.3186058>.

GERON, A. **AOS A OBRA: APRENDIZADO DE MAQUINA COM SCIKIT-LEARN & TENSORFLOW**. 1. ed. [S.l.]: Alta Books, 2019. ISBN 978-8550803814.

GOODFELLOW, I.; BENGIO, Y.; COURVILLE, A. **Deep Learning**. [S.l.]: MIT Press, 2016. <http://www.deeplearningbook.org>.

IBM. **IBM O que são redes neurais?** [S.l.]: IBM, 2022. <https://www.ibm.com/br-pt/topics/neural-networks>.

MAKINEN, M. et al. Inequalities in health care use and expenditures: empirical data from eight developing countries and countries in transition. **Bulletin of the world health organization, SciELO Public Health**, v. 78, n. 1, p. 55–65, 2000. Available from Internet: <https://pubmed.ncbi.nlm.nih.gov/10686733/>.

MONARD, M. C.; BARANAUSKAS, J. A. **Sistemas Inteligentes Fundamentos e Aplicações**. 1. ed. Barueri, SP: Manole Ltda, 2003. 89-114 p. (Conceitos sobre aprendizado de máquina). ISBN 85-204-168. Available from Internet: <https://dcm.ffclrp.usp.br/~augusto/publications/2003-sistemas-inteligentes-cap4.pdf>.

OKóLSKI, M. Spatial mobility from the perspective of the incomplete migration concept. **Central and Eastern European Migration Review**, v. 1, p. 11–35, 12/2012 2012. ISSN 2300-1682. Available from Internet: <http://www.ceemr.uw.edu.pl/vol-1-no-1-december-2012/articles/spatial-mobility-perspective-incomplete-migration-concept>.

OLIVEIRA, B. L. C. A. de et al. Evolução, distribuição e expansão dos cursos de medicina no brasil (1808-2018). **Trabalho, Educação e Saúde, SciELO Brasil**, v. 17, n. 1, 2019. Available from Internet: <https://doi.org/10.1590/1981-7746-sol00183>.

ONE. **One Hot Encoding in Machine Learning**. Geeks For Geeks, 2024. Available from Internet: <https://www.geeksforgeeks.org/ml-one-hot-encoding/>.

PADILHA, V. A.; CARVALHO, A. C. P. L. F. MineraÇÃo de dados em python. 2017. Available from Internet: <https://edisciplinas.usp.br/pluginfile.php/4052836/mod_resource/content/4/mineracaodadosbiologicos-parte3.pdf>.

POWELL, J. **A Librarian's Guide to Graphs, Data and the Semantic Web**. 225 Wyman Street, Waltham, MA 02451, USA: Elsevier Ltda, 2015.

RAHMAN, M. **What You Need to Know about Sparse Categorical Cross En- tropy**. medium, 2023. Available from Internet: <https://rmoklesur.medium.com/ what-you-need-to-know-about-sparse-categorical-cross-entropy-9f07497e3a6f>.

RIGATTI, S. J. Random forest. **Journal of Insurance Medicine**, v. 47, n. 1, p. 31–39, 01 2017. ISSN 0743-6661. Available from Internet: <https://doi.org/10.17849/insm-47-01-31-39.1>.

RóJ, J. Inequality in the distribution of healthcare human resources in poland. **Sustainability, Multidisciplinary Digital Publishing Instituteh**, v. 12, n. 5, 2020. Available from Internet: <https://doi.org/10.3390/su12052043>.

SEIXAS, P. H. D. et al. A circularidade dos médicos em cinco regiões de são paulo, brasil: padrões e fatores intervenientes. **Cadernos de saúde pública, II**, n. 35, 2019. Available from Internet: <https://doi.org/10.1590/0102-311X00135018>.

SEIXAS, P. H. D. et al. Physicians' circularity in health regions in brazil. **Revista Brasileira de Saúde Materno Infantil, I**, n. 17, 2017. Available from Internet: <https://doi.org/10.1590/1806-9304201700S100009>.

SMITH, C. **Decision Trees and Random Forests: A Visual Introduction for Beginners**. Blue Windmill Media, 2017. ISBN 9781549893759. Available from Internet: <https://books.google.com.br/books?id=Hi_CtAEACAAJ>.

TENSORFLOW. databricks, 2024. Available from Internet: <https://www.databricks.com/br/glossary/tensorflow-guide>.

UFRJ. **DICAS PARA A CONFIGURAÇÃO DE REDES NEURAIS**. 2024. Available from Internet: <http://www.nce.ufrj.br/labic/downloads/dicas_cfg_rna.pdf>.

VISHWAKARMA, N. **What is the Adam Optimizer?** Analytics Vidhya, 2024. Available from Internet: <https://www.analyticsvidhya.com/blog/2023/09/what-is-adam-optimizer/>.

WANG, J. et al. Urban human mobility: Data-driven modeling and prediction. **SIGKDD Explor. Newsl.**, Association for Computing Machinery, New York, NY, USA, v. 21, n. 1, p. 1–19, may 2019. ISSN 1931-0145. Available from Internet: <https://doi.org/10.1145/3331651.3331653>.

ZHANG, M.-L.; ZHOU, Z.-H. A review on multi-label learning algorithms. **Knowledge and Data Engineering, IEEE Transactions on**, v. 26, p. 1819–1837, 08 2014.

A ANEXO - CÓDIGOS IBGE ESTADOS

Tabela com a correspondência dos estados para seu código IBGE

Tabela A.1 - Códigos IBGE dos Estados Brasileiros

Estado	Código IBGE
Acre	12
Alagoas	27
Amapá	16
Amazonas	13
Bahia	29
Ceará	23
Distrito Federal	53
Espírito Santo	32
Goiás	52
Maranhão	21
Mato Grosso	51
Mato Grosso do Sul	50
Minas Gerais	31
Pará	15
Paraíba	25
Paraná	41
Pernambuco	26
Piauí	22
Rio de Janeiro	33
Rio Grande do Norte	24
Rio Grande do Sul	43
Rondônia	11
Roraima	14
Santa Catarina	42
São Paulo	35
Sergipe	28
Tocantins	17

Fonte: Da autora (2025).

B ANEXO - CENTRALIDADES BASE ESTUDANTES

Tabelas com os valores das centralidades para os estados do país

Tabela B.1 - Dados de Centralidade por Região e Ano

Estado	2000	2001	2002	2003	2004	2005	2006	2007	2008	2009	2010	2011	2012
SP	30632.0	16525.0	16525.0	11035.67	16525.0	11035.67	11035.67	11035.67	11035.67	11035.67	11035.67	11035.67	11035.67
RJ	17546.0	9972.5	9972.5	6830.67	9972.5	6830.67	6830.67	6830.67	6830.67	6830.67	6830.67	6830.67	6830.67
MG	0.0	13563.5	13563.5	9223.0	13563.5	9223.0	9223.0	9223.0	9223.0	9223.0	9223.0	9223.0	9223.0
ES	0.0	0.0	0.0	2094.0	0.0	2094.0	2094.0	2094.0	2094.0	2094.0	2094.0	2094.0	2094.0
AM	0.0	1.0	2572.0	2473.0	662.75	662.75	662.75	541.80	541.80	662.75	453.0	453.0	453.0
PA	0.0	0.0	2552.0	0.0	685.25	685.25	685.25	548.2	548.2	685.25	469.0	469.0	469.0
AC	0.0	0.0	0.0	201.0	78.25	78.25	78.25	62.80	62.80	78.25	52.5	52.5	52.5
RO	0.0	0.0	0.0	0.0	447.5	447.5	447.5	358.6	358.6	447.5	299.0	299.0	299.0
TO	0.0	0.0	0.0	0.0	295.75	295.75	295.75	237.0	237.0	295.75	198.33	198.33	198.33
RR	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	13.20	13.20	0.0	11.0
AP	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	49.5	49.5	49.5
SE	1.0	0.0	1652.0	0.0	0.0	327.5	0.0	255.13	281.86	255.13	255.13	255.13	255.13
MA	0.0	2588.0	0.0	0.0	859.67	491.33	588.6	370.38	422.71	370.38	370.38	370.38	370.38
CE	0.0	7826.0	7750.0	1.0	0.0	1461.0	1750.4	1120.5	1279.29	1120.5	1120.5	1120.5	1120.5
PE	0.0	0.0	0.0	0.0	1559.33	823.5	978.80	630.38	711.29	630.38	630.38	630.38	630.38
BA	0.0	0.0	0.0	0.0	2835.33	1466.83	1717.4	1107.13	1260.57	1107.13	1107.13	1107.13	1107.13
PB	0.0	0.0	0.0	0.0	1209.67	715.67	850.0	600.63	650.14	600.63	600.63	600.63	600.63
PI	0.0	0.0	0.0	0.0	0.0	482.5	578.4	364.5	416.0	364.5	364.5	364.5	364.5
AL	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	81.125	0.0	81.125	81.125	81.125
RN	0.0	0.0	0.0	0.0	0.0	0.0	0.0	264.25	300.71	264.25	264.25	264.25	264.25
RS	1.0	1.0	1.0	0.0	4423.5	4423.5	4423.5	4423.5	4423.5	4423.5	4423.5	4423.5	4423.5
PR	0.0	0.0	0.0	1.0	5120.5	5120.5	5120.5	5120.5	5120.5	5120.5	5120.5	5120.5	5120.5
SC	0.0	0.0	0.0	0.0	3332.0	3332.0	3332.0	3332.0	3332.0	3332.0	3332.0	3332.0	3332.0
MT	1.0	1.0	0.0	861.5	1631.0	1631.0	621.33	621.33	621.33	621.33	621.33	621.33	621.33
DF	0.0	0.0	0.0	1087.5	2149.0	2149.0	930.0	930.0	930.0	930.0	930.0	930.0	930.0
MS	0.0	0.0	0.0	434.0	0.0	0.0	323.0	323.0	323.0	323.0	323.0	323.0	323.0
GO	0.0	0.0	0.0	0.0	0.0	0.0	1603.0	1603.0	1603.0	1603.0	1603.0	1603.0	1603.0

Fonte: Da autora (2025).

Tabela B.2 Dados de Centralidade de Grau de Entrada por Estado e Ano

Estado	2000	2001	2002	2003	2004	2005	2006	2007	2008	2009	2010	2011	2012
SP	14735.0	7928.5	7928.5	5292.0	7928.5	5292.0	5292.0	5292.0	5292.0	5292.0	5292.0	5292.0	5292.0
RJ	9354.0	5754.5	5754.5	3992.33	5754.5	3992.33	3992.33	3992.33	3992.33	3992.33	3992.33	3992.33	3992.33
MG	0.0	6347.5	6347.5	4268.0	6347.5	4268.0	4268.0	4268.0	4268.0	4268.0	4268.0	4268.0	4268.0
ES	0.0	0.0	0.0	1039.33	0.0	1039.33	1039.33	1039.33	1039.33	1039.33	1039.33	1039.33	1039.33
AM	0.0	1.0	1334.0	1243.0	345.75	345.75	345.75	286.8	286.8	345.75	240.33	240.33	240.33
PA	0.0	0.0	1228.0	0.0	312.0	312.0	312.0	249.60	249.60	312.0	216.0	216.0	216.0
AC	0.0	0.0	0.0	94.0	32.75	32.75	32.75	26.20	26.20	32.75	22.0	22.0	22.0
RO	0.0	0.0	0.0	0.0	228.0	228.0	228.0	182.8	182.8	228.0	152.5	152.5	152.5
TO	0.0	0.0	0.0	0.0	166.25	166.25	166.25	133.4	133.4	166.25	111.67	111.67	111.67
RR	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	2.0	2.0	0.0	1.67
AP	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	22.0	22.0	22.0
SE	1.0	0.0	831.0	0.0	0.0	172.33	0.0	134.75	148.71	134.75	134.75	134.75	134.75
MA	0.0	1321.0	0.0	0.0	428.0	245.33	293.8	185.0	210.86	185.0	185.0	185.0	185.0
CE	0.0	3886.0	3870.0	1.0	0.0	682.33	818.40	517.5	590.86	517.5	517.5	517.5	517.5
PE	0.0	0.0	0.0	0.0	730.33	383.83	457.0	294.38	331.29	294.38	294.38	294.38	294.38
BA	0.0	0.0	0.0	0.0	1376.0	699.17	829.0	525.38	600.0	525.38	525.38	525.38	525.38
PB	0.0	0.0	0.0	0.0	697.67	453.17	536.2	397.75	419.0	397.75	397.75	397.75	397.75
PI	0.0	0.0	0.0	0.0	0.0	248.0	297.40	187.0	213.57	187.0	187.0	187.0	187.0
AL	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	25.75	0.0	25.75	25.75	25.75
RN	0.0	0.0	0.0	0.0	0.0	0.0	0.0	129.5	147.0	129.5	129.5	129.5	129.5
RS	1.0	1.0	1.0	0.0	2117.5	2117.5	2117.5	2117.5	2117.5	2117.5	2117.5	2117.5	2117.5
PR	0.0	0.0	0.0	1.0	2423.5	2423.5	2423.5	2423.5	2423.5	2423.5	2423.5	2423.5	2423.5
SC	0.0	0.0	0.0	0.0	1897.0	1897.0	1897.0	1897.0	1897.0	1897.0	1897.0	1897.0	1897.0
MT	1.0	1.0	0.0	440.0	807.0	807.0	328.67	328.67	328.67	328.67	328.67	328.67	328.67
DF	0.0	0.0	0.0	549.0	1083.0	1083.0	553.67	553.67	553.67	553.67	553.67	553.67	553.67
MS	0.0	0.0	0.0	202.5	0.0	0.0	166.67	166.67	166.67	166.67	166.67	166.67	166.67
GO	0.0	0.0	0.0	0.0	0.0	0.0	0.0	689.67	689.67	689.67	689.67	689.67	689.67

Fonte: Da autora (2025).

Tabela B.3 Dados de Centralidade de Grau de Saída por Estado e Ano

Estado	2000	2001	2002	2003	2004	2005	2006	2007	2008	2009	2010	2011	2012
SP	15897.0	8596.5	8596.5	5743.67	8596.5	5743.67	5743.67	5743.67	5743.67	5743.67	5743.67	5743.67	5743.67
RJ	8192.0	4218.0	4218.0	2838.33	4218.0	2838.33	2838.33	2838.33	2838.33	2838.33	2838.33	2838.33	2838.33
MG	0.0	7216.0	7216.0	4955.0	7216.0	4955.0	4955.0	4955.0	4955.0	4955.0	4955.0	4955.0	4955.0
ES	0.0	0.0	0.0	1054.67	0.0	1054.67	1054.67	1054.67	1054.67	1054.67	1054.67	1054.67	1054.67
AM	0.0	1.0	1238.0	1230.0	317.0	317.0	317.0	255.0	255.0	317.0	212.67	212.67	212.67
PA	0.0	0.0	1324.0	0.0	373.25	373.25	373.25	298.6	298.6	373.25	253.0	253.0	253.0
AC	0.0	0.0	0.0	107.0	45.5	45.5	45.5	36.6	36.6	45.5	30.5	30.5	30.5
RO	0.0	0.0	0.0	0.0	219.5	219.5	219.5	175.8	175.8	219.5	146.5	146.5	146.5
TO	0.0	0.0	0.0	0.0	129.5	129.5	129.5	103.6	103.6	129.5	86.67	86.67	86.67
RR	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	11.20	11.20	0.0	9.33
AP	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	27.5	27.5	27.5
SE	1.0	0.0	821.0	0.0	0.0	155.17	0.0	120.38	133.14	120.38	120.38	120.38	120.38
MA	0.0	1267.0	0.0	0.0	431.67	246.0	294.8	185.38	211.86	185.38	185.38	185.38	185.38
CE	0.0	3940.0	3880.0	1.0	0.0	778.67	932.0	603.0	688.43	603.0	603.0	603.0	603.0
PE	0.0	0.0	0.0	0.0	829.0	439.67	521.80	336.0	380.0	336.0	336.0	336.0	336.0
BA	0.0	0.0	0.0	0.0	1459.33	767.67	888.40	581.75	660.57	581.75	581.75	581.75	581.75
PB	0.0	0.0	0.0	0.0	512.0	262.5	313.8	202.88	231.14	202.88	202.88	202.88	202.88
PI	0.0	0.0	0.0	0.0	0.0	234.5	281.0	177.5	202.43	177.5	177.5	177.5	177.5
AL	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	55.38	0.0	55.38	55.38	55.38
RN	0.0	0.0	0.0	0.0	0.0	0.0	0.0	134.75	153.71	134.75	134.75	134.75	134.75
RS	1.0	1.0	1.0	0.0	2306.0	2306.0	2306.0	2306.0	2306.0	2306.0	2306.0	2306.0	2306.0
PR	0.0	0.0	0.0	1.0	2697.0	2697.0	2697.0	2697.0	2697.0	2697.0	2697.0	2697.0	2697.0
SC	0.0	0.0	0.0	0.0	1435.0	1435.0	1435.0	1435.0	1435.0	1435.0	1435.0	1435.0	1435.0
MT	1.0	1.0	0.0	421.5	824.0	824.0	292.67	292.67	292.67	292.67	292.67	292.67	292.67
DF	0.0	0.0	0.0	538.5	1066.0	1066.0	376.33	376.33	376.33	376.33	376.33	376.33	376.33
MS	0.0	0.0	0.0	231.5	0.0	0.0	156.33	156.33	156.33	156.33	156.33	156.33	156.33
GO	0.0	0.0	0.0	0.0	0.0	0.0	0.0	913.33	913.33	913.33	913.33	913.33	913.33

Fonte: Da autora (2025).

C ANEXO - CENTRALIDADE BASE PROFISSIONAIS

Tabelas com os valores das centralidade de grau para os estados do país

Tabela C.1 Dados de Centralidade por Região e Ano (2012-2020)

Estado	2012	2013	2014	2015	2016	2017	2018	2019	2020
MG	89761.33	89761.33	89761.33	89761.33	89761.33	89761.33	89761.33	89761.33	89761.33
SP	148656.67	148656.67	148656.67	148656.67	148656.67	148656.67	148656.67	148656.67	148656.67
RJ	50971.33	50971.33	50971.33	50971.33	50971.33	50971.33	50971.33	50971.33	50971.33
ES	26965.33	26965.33	26965.33	26965.33	26965.33	26965.33	26965.33	26965.33	26965.33
RS	80573.0	80573.0	80573.0	80573.0	80573.0	80573.0	80573.0	80573.0	80573.0
PR	66157.0	66157.0	66157.0	66157.0	66157.0	66157.0	66157.0	66157.0	66157.0
SC	53197.0	53197.0	53197.0	53197.0	53197.0	53197.0	53197.0	53197.0	53197.0
PA	12380.0	12380.0	12380.0	12380.0	12380.0	12380.0	12380.0	12380.0	12380.0
RO	6917.0	6917.0	6917.0	6917.0	6917.0	6917.0	6917.0	6917.0	6917.0
AM	3475.67	3475.67	3475.67	3475.67	3475.67	3475.67	3475.67	3475.67	3475.67
AP	1055.0	1055.0	1055.0	1055.0	1055.0	1055.0	1055.0	1055.0	1055.0
RR	1764.67	1764.67	1764.67	1764.67	1764.67	1764.67	1764.67	1764.67	1764.67
TO	3232.0	3232.0	3232.0	3232.0	3232.0	3232.0	3232.0	3232.0	3232.0
AC	2293.0	2293.0	2293.0	2293.0	2293.0	2293.0	2293.0	2293.0	2293.0
PB	14457.0	14457.0	14457.0	14457.0	14457.0	14457.0	14457.0	14457.0	14457.0
PE	34102.75	34102.75	34102.75	34102.75	34102.75	34102.75	34102.75	34102.75	34102.75
RN	8204.0	8204.0	8204.0	8204.0	8204.0	8204.0	8204.0	8204.0	8204.0
CE	38191.0	38191.0	38191.0	38191.0	38191.0	38191.0	38191.0	38191.0	38191.0
BA	39939.0	39939.0	39939.0	39939.0	39939.0	39939.0	39939.0	39939.0	39939.0
AL	5023.5	5023.5	5023.5	5023.5	5023.5	5023.5	5023.5	5023.5	5023.5
PI	6464.5	6464.5	6464.5	6464.5	6464.5	6464.5	6464.5	6464.5	6464.5
SE	5968.25	5968.25	5968.25	5968.25	5968.25	5968.25	5968.25	5968.25	5968.25
MA	12249.25	12249.25	12249.25	12249.25	12249.25	12249.25	12249.25	12249.25	12249.25
GO	41163.33	41163.33	41163.33	41163.33	41163.33	41163.33	41163.33	41163.33	41163.33
MT	6288.0	6288.0	6288.0	6288.0	6288.0	6288.0	6288.0	6288.0	6288.0
DF	5646.0	5646.0	5646.0	5646.0	5646.0	5646.0	5646.0	5646.0	5646.0
MS	6975.33	6975.33	6975.33	6975.33	6975.33	6975.33	6975.33	6975.33	6975.33

Fonte: Da autora (2025).